

DOCUMENT RESUME

ED 089 735

IR 000 447

AUTHOR Scheffler, F. L.; And Others
TITLE Feasibility of Computer Processing of Technical Information on the Design of Instructional Systems. Final Report for the Period 1 July 1972 through 31 March 1973.
INSTITUTION Air Force Human Resources Lab., Wright-Patterson AFB, Ohio. Advanced Systems Div.; Dayton Univ., Ohio. Research Inst.
REPORT NO AFHRL-TR-73-40
PUB DATE Jan 74
NOTE 110p.
EDRS PRICE MF-\$0.75 HC-\$5.40 PLUS POSTAGE
DESCRIPTORS Artificial Intelligence; Computer Oriented Programs; *Computers; Cost Effectiveness; Feasibility Studies; Information Needs; *Information Processing; Information Retrieval; Information Storage; *Information Systems; *Instructional Design; *Instructional Systems; *Man Machine Systems; On Line Systems; Use Studies

ABSTRACT

A feasibility study examined the capability of a computer-based system's handling of technical information pertinent to the design of instructional systems. Structured interviews were held to assess the information needs of both researchers and practitioners and an investigation was conducted of 10 computer-based information storage and retrieval systems which could serve as the basis for a technical information system. Question-answering computer systems relying on artificial intelligence concepts were reviewed and considered in terms of providing a prescriptive interactive system. It was concluded that practicable generalized semantic information retrieval systems would require further development. It was further decided that the design and implementation of a natural language automated information retrieval system encompassing the rudimentary features of a prescriptive system was feasible, both in terms of the technical information resources provided and its cost-effectiveness. Lastly, an on-line interactive retrieval system was designed and a model data base of approximately 500 literature-derived comprehensive abstracts was implemented and tested to confirm the feasibility of the system. (Author)

AFHRL-TR-73-40

**FEASIBILITY OF COMPUTER PROCESSING OF TECHNICAL
INFORMATION ON THE DESIGN OF INSTRUCTIONAL SYSTEMS**

By

F.L. Scheffler

F.J. DaPolito

R.L. McAdams

M.J. Gee

University of Dayton Research Institute

300 College Park Avenue

Dayton, Ohio 45469

**ADVANCED SYSTEMS DIVISION
Wright-Patterson Air Force Base, Ohio 45433**

January 1974

Final Report for the Period 1 July 1972 - 31 March 1973

Approved for public release; distribution unlimited.

LABORATORY

**AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235**

AIR FORCE



HUMAN

RESOURCES

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY.

ED 089733

IR 000447

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by the University of Dayton Research Institute, 300 College Park Ave., Dayton, Ohio 45469, under Contract F33615-72-C-2091, Project 1710, with the Advanced Systems Division, Air Force Human Resources Laboratory (AFSC), Wright-Patterson Air Force Base, Ohio 45433. Mr. Horace H. Valverde and Dr. Donald L. Thomas, Advanced Systems Division, shared the contract monitorship.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

GORDON A. ECKSTRAND, Director
Advanced Systems Division

Approved for publication.

HAROLD E. FISCHER, Colonel, USAF
Commander

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TR-73-40	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Feasibility of Computer Processing of Technical Information on the Design of Instructional Systems	5. TYPE OF REPORT & PERIOD COVERED Final Report (1 July 1972 - 31 March 1973)	
7. AUTHOR(s) F. L. Scheffler, F. J. DaPolito, R. L. McAdams, and M. J. Gee	6. PERFORMING ORG. REPORT NUMBER UDRI-TR-73-23	
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Dayton Research Institute 300 College Park Ave. Dayton, Ohio 45469	8. CONTRACT OR GRANT NUMBER(s) F33615-72-C-2091	
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62703F 1710 03 33	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Advanced Systems Division of Air Force Human Resources Laboratory (AFHRL/AST) Wright-Patterson Air Force Base, Ohio 45433	12. REPORT DATE January 1974	
	13. NUMBER OF PAGES 114	
	15. SECURITY CLASS. (of this report) UNCLASSIFIED	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution is unlimited		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
Document Retrieval Systems	Feasibility	Training Research In-
Information Retrieval Systems	Training Research	formation
On-line Information Systems	User Needs Study	ISD Information Sys-
Instructional Systems Design	Evaluation	tems (cont'd)
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)		
See reverse side		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

19. Prescriptive Guidance

20. A study was made of the feasibility of a computer-based system to handle technical information pertinent to the design and use of instructional systems. The study took into account the information needs of both researchers and practitioners. Current and projected information needs of researchers and practitioners were determined from structured interviews. Ten available computer-based information storage and retrieval systems which could serve as a basis for such a technical information system were investigated. Question-answering computer systems based on artificial intelligence concepts were reviewed and considered in terms of providing a prescriptive interactive system. It was concluded that practicable generalized semantic information retrieval systems would require further developments. The design and implementation of a natural language automated information retrieval system encompassing rudimentary features of a prescriptive system is feasible, both in terms of the technical information resource provided and its cost effectiveness. An on-line interactive retrieval system was designed, and a model data base of approximately 500 literature-derived comprehensive abstracts was implemented and tested to confirm the feasibility of the system.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

SUMMARY

PURPOSE

The purpose was to study the feasibility of a computer-based information storage and retrieval system for technical information on training research and instructional system design. Two basic types of information users need to be accommodated: the researcher and the practitioner.

APPROACH

A study was made of the current and projected information needs of both researchers and practitioners by conducting a series of structured interviews with typical potential users. Questions were asked to elicit characteristics of information need requirements. However, discussions in response to the questions were permitted to range to include each individual's expressions of information needs and his perception of a suitable information environment.

A survey was made of various available information storage and retrieval systems which could meet the apparent requirements. Ten different systems were studied from the standpoints of retrieval capability, input and updating requirements, data manipulation, accessibility by the user, hardware requirements and availability, and costs.

A review was made of question answering interactive retrieval systems based on artificial intelligence concepts. The purpose of this review was to project the likely state of development of such systems within a three- to five-year period with consideration of the need for consultive or prescriptive guidance in instructional system design.

RESULTS

The users' needs study revealed that four types of data bases could be considered: (1) a literature-derived data base, (2) a data base for interactive prescriptive guidance, (3) a data base listing experts in various phases of training and human performance, and (4) a data base consisting of a subject-index catalog of already designed and existing instructional systems. The need for cross-referencing between data bases was clearly seen as a desirable feature.

The investigation of available information storage and retrieval systems indicated that nearly all systems investigated were suitable for document retrieval. However, the techniques and software for an interactive prescriptive

or question-answering system is beyond the present state of the art. Of the information retrieval systems surveyed, the Avionics Central system operated by the Air Force Avionics Laboratory at Wright-Patterson AFB, Ohio is considered the most suitable. This system is a natural language interactive information retrieval system which provides for automatic indexing, file maintenance, and rapid retrieval of abstracts and related items.

The results of investigating question-answering systems based on artificial intelligence concepts show that recent theoretical break-throughs have brought the state of the art to the threshold of developing practicable generalized semantic information systems. Nevertheless, further significant advances are yet necessary before a validated and useful prescriptive system can be developed. Because of rapid developments in this field, an active current awareness should be maintained. The only type of prescriptive guidance system which appears practical within the three- to five-year time period is the "computerized handbook" concept or human-derived extractions from available literature and experts' knowledge.

CONCLUSIONS

The design and implementation of a natural language automated document retrieval system encompassing some rudimentary features of a prescriptive system is entirely feasible, both regarding the technical information resource provided and its cost effectiveness. An experimental information retrieval system to accommodate abstracts was designed using the USAF Avionics Central system. A model data base of approximately 500 literature-derived comprehensive abstracts was implemented and is currently being tested by personnel of the Air Force Human Resources Lab. The other data bases described above could be incorporated into the basic system concept. A more sophisticated prescriptive guidance interactive system could be developed currently as a computerized handbook with human-derived prescriptions. However a natural language conversational man-machine dialog question-answering system, while theoretically possible and probably ultimately achievable, requires further breakthroughs in semantic theory and artificial intelligence before such a system can be realized in practice.

PREFACE

This report was prepared by the University of Dayton Research Institute, Dayton, Ohio, under Air Force Contract F33615-72-C-2091. The work described herein was accomplished under Project 1710, "Training for Advanced Air Force Systems", Task 171003, "Training Implications of New Military Technology". The effort constituted a portion of the program of the Advanced Systems Division, Air Force Human Resources Laboratory. Dr. Ross L. Morgan is the Project Scientist. When the effort was initiated, Mr. Horace H. Valverde served both as Task Scientist and Contract Monitor. When Mr. Valverde retired in Dec. 72, Mr. Bertram W. Cream assumed the duties of Task Scientist and Dr. Donald L. Thomas assumed the duties of the Contract Monitor.

This is a final technical report and covers the work accomplished from 1 July 1972 through 31 March 1973.

The authors acknowledge the assistance and cooperation of other organizations and personnel. In particular we are grateful to the Avionics Laboratory for permitting us to establish a model data base with the Avionics Central software system, to the Foreign Technology Division for converting MT/ST data to computer-readable form, and to the Office for Computing Activities of the University of Dayton for reformatting data into the proper input format for the model data base, and to the Applied Science Associates for cooperation in providing suitable abstracts.

Mr. Robert R. Roalef, Air Force Avionics Laboratory
Mr. Frank G. Jones, Air Force Avionics Laboratory
Mr. William R. Mace, Air Force Foreign Technology Division
Mr. Jack A. Pugh, University of Dayton
Mr. Sanford P. Schumacher, Applied Science Associates
Mrs. Jacqueline F. March, University of Dayton Research Institute
Miss Deborah M. Rekart, University of Dayton

TABLE OF CONTENTS

SECTION	TITLE	PAGE
1	INTRODUCTION	7
2	STUDY OF USERS' NEEDS	9
	2.1 Methodology	9
	2.2 Types of Users	10
	2.3 Results of Interviews	11
	2.3.1 Areas of Work and Interest	11
	2.3.2 Current Sources of Information	12
	2.3.3 Identification of Information Needs	13
	2.3.4 Examples of Situations Involving Information Needs	14
	2.3.5 Concept of an Information System	15
	2.3.6 Use of Information, Response Time, and Current Awareness Needs	18
	2.4 Summary of Users' Needs	19
3	INVESTIGATION OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS	20
	3.1 Review of Currently Available Systems	20
	3.2 Criteria for Selection of an IS&R System	21
	3.3 Results of the Information Systems Survey	22
	3.4 Recommendation	23
	3.5 Projection of the State of the Art of Computer-Based IS&R Systems in 3-5 Years	24
	3.5.1 Storage	24
	3.5.2 Input Devices	25
	3.5.3 Output Devices	25
	3.5.4 Data Processing	26

TABLE OF CONTENTS

SECTION	TITLE	PAGE
4	NATURAL LANGUAGE QUESTION-ANSWERING SYSTEMS	27
5	ESTABLISHMENT AND IMPLEMENTATION OF A MODEL DATA BASE OF COMPREHENSIVE ABSTRACTS	31
	5.1 System Design	31
	5.2 Retrieval from the Model Data Base	40
	5.2.1 Computer-user-Dialog	40
	5.2.2 Search Procedures	42
	5.2.3 Display	45
	5.3 Indexing, Updating, Maintenance	46
	5.4 Costs	48
REFERENCES		50
APPENDIX A	DESCRIPTION AND ANALYSIS OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS INVESTIGATED	51
	Document Processing System	52
	Mead Data Central	56
	Aerospace Materials Information Center	61
	BASIS-70	65
	Remote Information Query System (RIQS)	69
	DIALOG	74
	ORBIT II	77
	NASIS	80
	Query/Update	85
	STAIRS	88
APPENDIX B	USERS' MANUAL FOR THE MODEL DATA BASE OF COMPREHENSIVE ABSTRACTS USING AVIONICS CENTRAL	92

TABLE OF CONTENTS

SECTION	TITLE	PAGE
B-1	Introduction	94
	B-1.1 Description	94
	B-1.2 Term Definitions	94
B-2	Operating Instructions	97
	B-2.1 Starting Up	97
	B-2.2 Communicating	98
	B-2.3 Requesting a Search	98
	B-2.4 Modifying a Search	103
	B-2.5 Displaying Search Results	104

SECTION I

INTRODUCTION

An information storage and retrieval system in the area of training research and instructional system design potentially could be a valuable tool in aiding researchers and practitioners in performing their work. A feasibility study was conducted to determine: (1) the functions and characteristics that should be incorporated in the design of such an information system, (2) the availability and capabilities of information storage and retrieval systems already developed, (3) the approximate costs of maintaining and operating such an information system, (4) procedures required for maintaining, updating, editing, and retrieving information from the system, and (5) the relevant anticipated developments in information system technology in the next three to five years.

An effective information system could serve as a valuable resource for researchers and practitioners concerned with training research and the development of courses of instruction, training aids, and training devices. Several benefits can be anticipated from having an information system. The information would serve as a central source of technical information and data. Thus, better utilization of technical information resources should result. Depending on the data bases incorporated and the means provided for indexing and retrieval, the accessibility of available technical information should be greatly enhanced.

If a prescriptive type information system could be developed, effective guidance could be provided to personnel involved with designing all facets of an instructional system. An interactive prescriptive system would be able to provide certain facts and principles which the instructional system designer could apply in developing training courses. With such a system, alternatives for presenting training material could be explored by the instructional system designer so that he would have a basis upon which to select more effective modes of training.

A literature-derived data base of abstracts would permit the researcher to identify and review pertinent literature concerning the topic of interest. It would provide an effective aid for identifying areas which are important for instructional systems design but in which little research has been accomplished. The information system would increase the probability that all relevant research is reviewed, would reduce the time and effort required to accomplish the review, and would improve accessibility to the desired information by providing alternative searching methods.

The project was conducted in four phases to accomplish the purpose of the feasibility study. These phases are as follows:

1. Study of Users' Needs
2. Survey of ten information systems
3. Review of the state of the art of question-answering systems
4. Design, establishment, and testing of a model data base.

Each of these phases is treated in a separate section in the balance of the report.

SECTION 2

'STUDY OF USERS' NEEDS

2.1 METHODOLOGY

A study of the needs of potential users of a technical information system on training and instructional system design was conducted. The methodology which was used had proven to be highly successful in an earlier project.¹ This methodology consists of a structured interview technique involving small groups of potential users.

Individual and small group interviews were conducted with research personnel of the Advanced Systems Division of the Air Force Human Resources Laboratory (AFHRL) and with a highly experienced designer of instructional systems from the Air Training Command Headquarters at Randolph Air Force Base. The purpose of the interviews was to solicit input from potential users of an information system on training, human learning and instructional systems. Small group interviews have the advantage of providing the opportunity for interactions between the interviewer and the interviewee as well as among the interviewees. Such interactions resulted in new insights and new approaches which were valuable in determining the feasibility and design characteristics of a suitable information system. These interviews were conducted by an information specialist and by a psychologist knowledgeable in instructional technology. In conducting the interviews, suggestions were made about functions which an information system could perform. Feedback was received in response to these suggestions. Also, questions were asked concerning present methodologies used in training research and designing instructional systems. It was important to identify information which would have been useful if it had been available.

A structured list of questions was presented to the potential users to provide guidelines for deriving appropriate information and to maintain uniformity of interviews among the various personnel interviewed. All interview sessions were tape recorded.

1. F.L. Scheffler, A.T. Dodson, T.R. Featheringham, and J.P. Webb, A Survey of the Technical Information Environment of the Air Force Flight Dynamics Laboratory, UDRI-TR-69-29, University of Dayton Research Institute, Dayton, Ohio, October 1969.

2.3 RESULTS OF THE INTERVIEWS

The interviews revealed a number of interesting facts, both regarding the opinions of AFHRL personnel on currently available information resources and on their attitudes concerning the information environments they envisioned as being desirable.

A set of eight questions provided the basic structure of the interviews. At this point, the concern was to determine the users' needs without necessarily relating these needs to an information system. The questions served as starting points for discussion. These questions were as follows:

1. What is the area of your work and interests?
2. What current sources of information do you use when you have a n information need?
3. Can you identify information needs which must be satisfied in order to conduct your research?
4. Give an example of a recent experience which required access to information and indicate how this need was met.
5. What would you like to have for an information system assuming no technical or economic restraints?
6. What use would or could you make of information provided? Is hardcopy required?
7. What approximate turnaround time is required?
8. Do you have a current awareness need for information?

A summary of the responses to these questions and the resulting discussions is given in the following paragraphs.

2.3.1 Areas of Work and Interests

The areas of work and interests described by the researchers interviewed were largely concerned with media and the effective use of media in training situations. The concept "media" in a training context concerns everything from textbooks and handbooks to audio-visual aids and rather sophisticated simulators and mock-ups. Task analysis, measurement criteria and procedures, and performance evaluation were also indicated as

important areas of interest. A real need exists for reliable criteria for measuring and evaluating training effectiveness.

Practitioners are interested in the actual development of training courses for specific training situations. They are less concerned with the reasons for the effectiveness of training techniques; rather they need to know techniques which will work for their particular training situation. The reliability of training techniques is of concern to the practitioner. The practitioner needs to be assured that the procedures he uses in designing a training course will be able to develop effectively the skills needed by the trainees.

2.3.2 Current Sources of Information

Some interesting observations were made regarding current sources of information. AFHRL personnel expressed the opinion that numerous deficiencies exist in current literature sources. One severe problem is the lack of standardization and constancy of terminology. Often new phrases or terms are created to refer to already existing concepts described by other terminology. Another serious problem is that author-generated or information center-generated abstracts of technical literature are often inadequate. Thus, to screen articles or reports of possible interest, the researcher cannot depend on the abstracts alone, but he must usually read a major portion of the document, often, only then, to discover that the article is of no value. If more pertinent information and data were provided in a convenient format, the user could determine the probable relevance of the documents with far greater efficiency. With current literature sources and formats, the frustration factor often becomes so great that the researcher frequently resorts to his own methods with inadequate information. AFHRL personnel are literature conscious and would make more extensive use of literature if the literature itself were formatted and described more adequately. Retrieval of technical reports or journal articles is difficult because keywords or subject index terms provided in manual indexes are usually too imprecise or general.

Implications for an information system with a literature-derived data base were discussed. Discussions indicated that a retrieval system should have the following characteristics: (1) the categorization of articles into laboratory studies, field studies and correlational studies; (2) the inclusion of functional relationships of the dependent and independent variables studied in the experiment; (3) the description of the conditions of the experiment (type of subjects, training environment, instruments used, media used, methodology employed etc.); and (4) a statement of the conclusions derived from the study. It would be desirable for a "quality" or "validity" rating to be provided for each document; however, it was recognized that such a rating would be subjective and would not be suitable for all users, since each user has a different need and viewpoint.

A valid taxonomic scheme could prove useful in organizing technical literature in training and instructional system design, but the probability of developing a sound taxonomy of training and instructional concepts which would be acceptable among the experts in the field seems very low.

Both researchers and practitioners tended to rely on current information sources in the following order of importance:

- (1) Personal knowledge and experience
- (2) Contacts with colleagues and personnel in appropriate disciplines
- (3) Personal library collection of training documents
- (4) Government Technical Reports
- (5) Key scientific and technical journals
- (6) Selective dissemination of information (SDI) bibliographies and abstract journals, such as Research in Education
- (7) Air Force manuals, state-of-the-art reviews, handbooks, technical books, and bibliographies

2.3.3 Identification of Information Needs

The question concerning identification of information needs which must be satisfied to perform the work of AFHRL personnel was difficult to answer. Both practitioners and researchers indicated that information needs vary widely depending on the specific assignments on which they were working.

The identification of information needs is basically an analytical problem, i.e., to determine what constitutes the essential information one needs to answer a question. If one could determine the basic information units or elements he needs, the problem then arises of how to acquire reliable data and information relating to these needs and integrating the informational units to a proper answer. One primary problem is to correlate available theoretical and experimental studies and results to applied practical problems in a reliable fashion.

Air Force training requirements are usually complex and consist of many task-oriented steps. The trainees have a very diverse range of skills and knowledge prior to commencing a training program. At the Advanced Systems Division, researchers are concerned with future long-range efforts of a multi-disciplinary nature. Therefore, information needs are

difficult to state and can cover a wide range. An item of information which was of little or no interest previously may suddenly become very significant. Because of the dynamic nature of training information needs, an information storage and retrieval system must be highly flexible in order to accommodate these needs.

2.3.4 Examples of Situations Involving Information Needs

An example of a response to a query from an external Air Force organization was given. A requester from the Tactical Air Command (TAC) wanted information regarding the use of a tachistoscope for aircraft recognition training. He had a very practical need, since he wanted to prepare an instructional program for classes he was going to teach in aircraft recognition. Specifically, he needed to know whether the tachistoscope would be an effective training medium and what conditions should be used, e.g., how many exposures, exposure durations, views of the aircraft, partial views such as tail sections, and views under degraded conditions. This problem was assigned to a researcher from the Air Force Human Resources Laboratory. He consulted the Division Library and found World War II reports on aircraft recognition. From a Defense Documentation Center (DDC) publication he found one very good recent report. The researcher, in essence, obtained a bibliography. From this bibliography he digested the material and provided a written summary of his interpretation of the findings along with the particularly pertinent technical report obtained from DDC to the original requester.

Another example dealt with planning of research in the area of adjunct programming. Adjunct programming concerns the presentation of test questions in conjunction with written instructional materials. Specific questions concern such factors as the frequency of testing; the types of tests to be administered (multiple choice, short answer, essay); and whether tests should be given before, during, or after the presentation of the instructional material.

A problem requiring comprehensive relevant information on the Systems Approach to Training (SAT) or Instructional System Design (ISD) was cited. In acquiring relevant information, no existing reference list was found to be available. Subsequently, literature indexes and services were consulted and reports obtained. As each relevant document was identified, acquired, and studied, the bibliography provided therein was then used to provide pathways to additional relevant materials. The authors of relevant documents were noted and searches for other documents and articles by those authors were sought.

The practitioner, in the Air Training Command (ATC), a highly experienced person in applied training, suggested two types of queries which must be handled. The first type is a question concerning general procedures. For example, one requester wanted to know how he could most effectively establish a five-day middle management course in management procedures and policies. In answering a query of this nature, the practitioner first attempted to find out what specific training courses were available on the topic. Had they been available, a listing of any such courses would have been obtained and examined. If such a course had been discovered in this way, the relevant course descriptions would have been obtained directly. Other suggestions for administering the course might also have been obtained.

The second type of query indicated by the practitioner is of a more specific nature. For example, he may be asked how important color is in training films. To answer this type of query, he would search technical reports and state-of-the-art reviews. The bibliographies contained therein often guide him to more appropriate specific documents. He provides the requester with appropriate document references and pertinent excerpts from documents on hand.

Discussions were held concerning the information activities in which the Advanced Systems Division of AFHRL is currently engaged, and future information requirements. Although queries from outside organizations are received, the more prevalent need is in planning research activities on more generalized training/instructional problems which the Division has predicted or detected or which have been suggested by some organization. However, with a good information system there could be considerably more information activities involving responses to queries. The questions exist, but there is no focal point to which training-related questions can be directed. It is expected that the quantity of such questions would increase greatly if a good computer-based technical information system were available.

2.3.5 Concept of an Information System

To focus on the desirability and feasibility of a computer-based information storage and retrieval system, it was important to derive from potential users their concept of an information system with specific consideration of the functions to be fulfilled by the system. Some responses to this question were quite specific regarding particular access or retrieval points which should be provided. Other responses were more concerned with the levels of sophistication which would be possible, and the significance of such levels of sophistication to the user. The user's ability to interact effectively with the system in accomplishing his information seeking tasks was emphasized.

Specific retrieval parameters which should be provided include: subject area; author; title; independent and dependent variables studied; date; corporate author; sponsoring organization; journal source; type of study (lab experiment, field study, correlation study); instrumentation, stimuli, media or hardware used; type of subjects; methodology or experimental design; modality; training environment; overall quality rating of the study. It would be highly desirable for the system to be able to display on request the bibliography or list of references contained in the retrieved documents. In addition to retrieval, it was suggested that the ideal system would also manipulate stored information such that areas of needed training research would be pinpointed, contradictory results would be indicated, and needs for additional research would be suggested.

The practitioner would like to see a computer-based information system with a data base both for applicable research literature and for existing instructional programs on various subjects. The system should be updatable so that invalid material could be deleted and replaced and so that new programs and modifications of existing programs in the data base could be added. The following descriptive and/or retrieval information should be supplied with a narrative description of the instructional system.

- (1) Date that the course was developed.
- (2) Time required to complete the course; total instructional hours.
- (3) Objectives of the training course; include aircraft/weapon system if appropriate.
- (4) Location (responsible military organization) of training program.
- (5) Name of the organization which developed the training course.
- (6) Equipment required.
- (7) Trainee characteristics (previous training, average military rank, educational level, etc.).
- (8) Evaluation of effectiveness.

It was suggested that a correlation of job specialty codes with training courses would be useful. AF Manual 39-1 and AF Manual 36-1 were cited as examples. A typical question which might be addressed to

the system would be: "what self-instruction courses are available for the maintenance of certain electronics equipment?"

In discussing the more general characteristics of information systems, three levels of sophistication were indicated. The most sophisticated system would consist of an interactive on-line system which could be used by the practitioner in the field to assist him in an instructional system development problem. The interaction would occur such that the system would lead him like a consultant in natural or near-natural English to refine his query until a well-defined search strategy had been formulated. In many respects, the interactive program would perform similarly to a Computer-Aided Instruction (CAI) program. The computer could suggest additional keywords and phrases to the user which he could select at his option to provide as precise a search strategy as appropriate. The search of the system would result in prescriptive guidelines to the user. For example, in posing the problem of aircraft recognition, the ideal prescription would provide detailed advice such as: "Show seven shapes for 0.2 seconds at five-second intervals". Whatever the prescriptive statements or principles, an indication of the level of confidence for the statement should be given, and appropriate source materials from which the statement is derived should be provided. Another feature of this most-sophisticated level would be the ability to manipulate specific data and information such that new concepts could be generated.

The second level of sophistication would permit an expert in instructional systems to obtain guidance to aid him in solving instructional problems by retrieving prescriptive statements with considerably less prompting from the computer in formulating his query or search strategy. The results of the search would be basically the same with either system, i.e., principles or prescriptive statements. Preferably, an indicated confidence level and references to bibliographic source material would be provided. One way of conceptualizing this second level of sophistication is as a "computerized handbook" of instructional technology. A handbook is basically a distillation and extraction of facts and principles derived from extensive research and application-oriented studies. The computer-based prescriptive system would present the user with this type of information in response to queries of the system.

The third level of sophistication would be similar to document storage and retrieval systems in widespread use today. With such a system, an expert in instructional research and development would retrieve bibliographic references in response to his search. For example, if the user requested all references on "PROFICIENCY MEASUREMENT", the system might respond, "250 documents satisfy your request." By successive interactions, additional search restrictions can be applied by requiring additional search terms and/or by limiting the output by date, by type of document, or by some other specifiable retrieval parameters until a reasonable number

of retrievals is obtained. Display of document titles, authors, abstracts, and other records would be possible. The end result of a search would be a bibliographic listing. Detailed study on the part of the user would require him to refer to the original document.

2.3.6 Use of Information, Response Time, and Current Awareness Needs

The use of information generated by an information system may vary depending on whether the user is a practitioner or a researcher. The practitioner generally would use information directly, to apply it to designing instructional courses. Both researchers and practitioners indicated that results from an information system would be used to assist in consulting.

Researchers would use the information system to identify areas needing research, to obtain research information related to a topic or project, and to learn of research procedures used on similar projects. In performing a search of the data base, the researcher may be able to confirm the need for a research project by the fact that the system did not reveal material on the topic being considered. The search results may show that research in the area of consideration is uncoordinated and isolated. Such results from the search of the system could offer guidelines to the researcher in defining his project, so that past research can be utilized and gaps in research can be filled in a systematic manner.

The response time required also would vary depending on the use to be made of the information. A practitioner designing a course of instruction needs immediate response. A researcher or practitioner in a consulting role needs a response preferably of less than one day. Researchers developing and planning new areas of research can tolerate much longer response times, perhaps as long as two weeks. However, in any case, the ability to search an information system interactively was strongly indicated as being highly desirable. Interactive capability would permit negotiation of the search request with the system until appropriate results were obtained. This capability would be especially useful in preventing the researcher from pursuing literature searching which ultimately would prove to be of no value.

Current awareness needs, while important, do not appear to be crucial in terms of an information system. Selective Dissemination of Information (SDI) as a spin-off from updating an information system on topics related to training research would be a desirable, but not essential, function. Both researchers and practitioners feel that sources they now employ -- reading of selected journals, receipt of current awareness announcements, and daily interaction with colleagues and others -- are generally sufficient.

2.4 SUMMARY OF USERS' NEEDS

Potential users of an information system related to instructional system design and training research need to have two basic requirements satisfied. Specifically, the practitioner must have access to reliable factual prescriptive-type data and information to assist in designing instructional systems for specific training situations. Researchers need to become familiar with the state of the art for specific problems and for planning research projects. An information system to serve as an easily accessible comprehensive resource for consultation would be highly desirable.

When asked to envision an information system with no technological or economic restraints, the users suggested three levels of sophistication. At the highest level, a highly interactive system would guide the information seeker in obtaining relevant facts and data. The ability to manipulate and integrate basic information and data to develop and present prescriptive guidelines would be useful. At an intermediate level, a "computerized handbook" concept was described. This concept would provide basic facts and principles with appropriate cross references to original sources. The most basic level would consist of a document retrieval system. A number of operational document retrieval systems are currently in use. The basic level would identify pertinent documents and would provide information (such as references, abstracts, and the experimental procedures used) for each document.

The need for an interactive system was clearly indicated, even for situations not requiring immediate response, since search negotiation capabilities are needed to ensure relevance of information retrieved. Comprehensive accurate descriptions of all the pertinent factors of concern in reported research should be provided, not only to permit the researcher to evaluate the retrieved information but also to allow him to specify certain conditions as retrieval requirements.

The incorporation of multiple data bases within the same information system is a definite possibility. Four possible data bases were identified for consideration: a literature-based data base, a data base for interactive prescriptive guidance, a data base listing experts in various phases of training research and instructional design, and a data base consisting of a subject-indexed catalog of already designed and existent instructional systems. Cross-referencing between data bases would be desirable. It would be essential for the prescriptive guidance system to cross-reference the literature sources from which prescriptions were derived. Because of limitations of scope and time, attention was concentrated on the feasibility of the prescriptive guidance data base and the literature-based retrieval data base.

SECTION 3

INVESTIGATION OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS

3.1 REVIEW OF CURRENTLY AVAILABLE SYSTEMS

To establish the feasibility of a computer-based information storage and retrieval (IS&R) system for technical information applicable to training research and the design and use of instructional systems, the state of the art of currently available IS&R systems was determined. Early in the study, a plan was adopted to explore thoroughly the capabilities of existing IS&R systems and to determine the suitability of these systems for designing and maintaining an information system appropriate for instructional system design and training research. The use of an existing system was considered preferable because of the high cost of developing software for an entirely new system.

Many individual IS&R systems are in existence. Each system has its own characteristics and idiosyncrasies regarding file structures, mode of addressing files and data elements within files, logical operators, and updating procedures. Certain systems are designed to be efficient from the standpoint of updating, others are more oriented toward rapid response during the retrieval operation. Some systems are geared for arithmetic logical operators and manipulation of data, whereas others provide for simple Boolean logical statements for retrieval. Information storage and retrieval systems can be classified as batch mode or on-line systems. They can be categorized as text processing automatic indexing systems or assigned keyword systems. Various configurations are possible, depending on the desired characteristics of the specific IS&R system application.

Based on a brief review of available IS&R systems, ten were selected for study. The selection was made on the basis of their being representative of the various IS&R techniques currently in use, in terms of the computer file handling techniques and user interaction with the system. Also, consideration was given to general purpose systems rather than systems designed primarily for very specialized applications. The systems selected for study are as follows:

<u>Name</u>	<u>Vendor/User</u>
Document Processing System	IBM/Foreign Technology Division, USAF
Mead Data Cnetral	Mead Corporation/Air Force Avionics Laboratory (Avionics Central)
Aerospace Materials Information Center	Air Force Materials Laboratory
BASIS-70	Battelle Memorial Institute
Remote Information Query System (RIQS)	Northwestern University
DIALOG	Lockheed Information Systems
ORBIT II	System Development Corporation
NASIS	Neoterics, Inc. /NASA
Query/Update	Control Data Corporation
STAIRS	IBM/State University of New York

A detailed description and analysis of these systems is presented in Appendix A.

3.2 CRITERIA FOR SELECTION OF AN IS&R SYSTEM

Based on a review of the users' needs study and the costs and capabilities of the various available systems, criteria were established for the selection of a specific IS&R system. Selection of a system was based on the following criteria:

- (1) An interactive on-line system was considered to be of prime importance.
- (2) The system should be very easy to access using a computer terminal for users not intimately familiar with programming or computer operations.
- (3) Rapid response time and good interactive capabilities should be provided.

- (4) A text processing automatic indexing system would be desirable.
- (5) The feasibility of making the system available to AFHRL personnel must be seriously considered.
- (6) The economic factors involved, including telephone line charges, hardware required to operate the system, storage requirements, and operational costs should be reasonable.

3.3 RESULTS OF THE INFORMATION SYSTEMS SURVEY

Review of the IS&R systems studied showed that any one of the ten systems could be used, but to achieve the most suitable information system factors such as the various system features, costs, and availability to AFHRL were carefully considered in selecting a specific system. The information system characteristics were reviewed in terms of data base creation, user interaction, special features, hardware and software requirements, costs, and advantages and disadvantages. The advanced state of development of most of the systems for document retrieval operations indicated that they met the criteria established for selection. Thus the final selection required close scrutiny of trade-offs in system performance, costs and availability to AFHRL, practitioners, and other users.

The Avionics Central system operated by the Avionics Laboratory with Mead Data Central^{*} software satisfies the criteria established. The information that would be desirable to put into a data base can be composed of the full text of the document in machine-readable form. The text processing automatic indexing feature provides that every word of every document (except for defined common words such as THE, OR, BUT, IT, etc.) becomes a retrieval parameter. Thus, a powerful search capability is available to the user. He can retrieve by word phrases as well as by single words and logical combinations thereof. The internal file structure of this system is a completely controlled indexed sequential structure wherein the data content of the user-defined data base is stored in both a sequential (serial-indexed) and an inverted (word-indexed) mode. The structure is oriented toward terminal inquiry response, and the user has no need for physical control over the internal structure.

When adding, deleting, or changing documents, the update programs essentially completely rebuild the system. Therefore, it is sometimes more efficient to treat massive files as separate data bases rather than to merge

^{*}The Avionics Central System operates primarily with the Mead Data Central Software. However, additional capabilities have been added to the system by Avionics Laboratory personnel.

them with smaller, more rapidly changing files. Additional segments can be added to existing files at any time, whenever it is determined that they are needed. This system is open-ended and extremely flexible.

The data base is created on random access disk storage by a set of system programs. Whenever a new data set is created or updates to the files are made, these files are copied onto magnetic tapes in order to provide security and stability for the user.

The user addresses the system in an on-line manner using a CC-30 terminal or a Teletype by an "access key" or identification code which determines those data bases and files of the accessible data base which he may use.

The tutorial capability provided by this system is excellent and will enable any user to become familiar with the commands quickly. One of the special features is the ability of the system to "depluralize" words that it is given. Thus, LESSON and LESSONS are treated as exactly equivalent forms. Also, an equivalence table can be built in by the user to accommodate synonymy of terms and spelling variations of the same term, e.g., COLOR and COLOUR.

Because of the text processing automatic indexing capabilities of the Avionics Central System, it is suitable not only for a literature-based document retrieval system, but it would also accommodate data bases of the computerized handbook, the reference file of experts, and a compilation of previously-developed instructional systems. Appropriate data base design and retrieval characteristics could be incorporated readily.

Although the system is fully operational, the software is constantly undergoing further development. Consequently, the capabilities of the system are constantly increasing to meet new user needs.

3.4 RECOMMENDATION

The Avionics Central system of the Avionics Laboratory at Wright-Patterson Air Force Base is recommended as the computer-based system for the efficient storage and retrieval of technical information applicable to training research and the design and use of Air Force instructional systems. The Avionics Laboratory has a policy of furnishing information storage and retrieval capabilities to various users beyond its own group on WPAFB. Therefore, one of the significant advantages is that proven software and hardware are already in operation by competent people who are willing to make the system available to other agencies. In this way cost sharing is accomplished, so that smaller users can have the benefits of a comprehensive system at a reasonable cost.

The Avionics Central System is available presently from 0745-1045 and 1315-1600 EST. The system is accessible through a dial-up terminal on local phone lines. Both 10 character per second and 30 character per second transmission rates are available. In testing this system, we found response times to be less than 30 seconds, and usually in the 5-10 second range. The local supplier has provided a very comprehensive system with on-line retrieval and batch updating capabilities. The documentation, including a section on definition of terms and the provision of examples, is well done.

This recommendation is based on information that is currently available. However, if requirements and circumstances change, consideration should be given to evaluating again Control Data's Query/Update, IBM's STAIRS, and Battelle's BASIS 70.

3.5 ANTICIPATED DEVELOPMENTS IN THE STATE OF THE ART OF COMPUTER-BASED IS&R SYSTEMS IN 3-5 YEARS

3.5.1 Storage

The most significant advance in the state of the art in three to five years will be the reduction in the cost per byte of mass random access storage. Devices that will make this possible include the TERA-BIT* 20-inch tape memory system now marketed by Ampex and the SCROLL 2-inch tape memory system which is under development at Control Data Corporation. Bubble memories and laser driven memories will also compete for their share of the mass storage market. In addition, devices commonly known as disc storage systems will continue to improve slightly in cost/performance and in storage capacity; however, the limit of such devices probably will soon be reached in terms of storage density as well as the cost/performance ratio. Until all of the bugs are eliminated from the 20-inch and 2-inch magnetic tape systems mentioned above, it is anticipated that the standard 1/2 inch tape storage systems will continue to be used with densities of 2300 bits per inch (bpi) becoming more common.

Within the next five years the cost/performance ratio of mass storage devices is expected to decrease by a factor somewhere between 10 and 100. By the late 1970's the increased use of Metal Oxide Semiconductor, Large Scale Integration (MOS, LSI) will continue to decrease the cost for main memory. It is anticipated that by the late 1970's, main memories will be of the 2-4 million byte size for a medium-priced computing system.

* This product name comes from the term TERA (for trillion), used to identify the capacity of the system.

3.5.2 Input Devices

The cost of data entry also will decrease significantly by the late 1970's, as the cost for video display devices which will be connected to disc or tape storage systems decreases. Continued research on cathode ray tubes (CRT's) and the new "plasma" display device will result in some improvement. However, the decrease in cost of the LSI logic and memory will reduce the cost of data entry most dramatically.

For information storage and retrieval of literature and other published reports, an alternative data entry procedure has already been developed and will be much more commonly practiced in 3-5 years. This technique involves the preparation of the article prior to publication in machine-readable form. As the various editing processes proceed for a given publication, this machine-readable code will also be edited and a machine-processable form of the final journal article will be distributed to appropriate centers for computer storage and retrieval. This technique will eliminate the most costly aspect of preparing data bases of journal articles and technical reports which exists at the present time, namely, the need for converting hardcopy printed form to machine-readable form. Updates, supplementary data and information, and editing corrections to the data bases will be made by the input devices already mentioned. In addition, it is anticipated that within the next five years audio response input devices may be perfected sufficiently for large-scale use. The use of such devices would completely eliminate the need for keying information for input.

3.5.3 Output Devices

The anticipated continually decreasing costs of video display devices for input are also significant with regard to the display of output information. Within the last two years, the costs of video display devices have decreased by a factor of three. During the next five years, it is anticipated that these costs will continue to decrease. More importantly, however, the cost/performance ratio will decrease even more. "Intelligent terminals" will be available at the price of current standard terminals, and the capability for display in color, even though more expensive, will be much more available and more common than at present. Greater strides in audio response on output will occur, although much less emphasis will be placed on audio responses than on video and hardcopy output. Computer Output Microforms (COM) will continue to be an important item, simply because of the speed with which these forms can be generated and their usefulness in storage space efficiency and ease of retrieval. As far as the projected information storage and retrieval requirements of the AFHRL are concerned, the video terminal with an optional hardcopy capability is expected to be the most important output device, and these devices are expected to decrease in price by a factor of 2 in the next three to five years.

3.5.4 Data Processing

As has already been mentioned, the decrease in costs for logic and integrated circuits used in memory will play a significant role in decreasing the cost of computing systems in the late 70's. More important than costs however, is the fact that new file structures and file access procedures will be available in packaged form which will allow the use of content addressable memory. With such a package, memory can be addressed by content rather than by location. The impact this development will have on information storage and retrieval is of vital significance. Information contained within main memory can be addressed immediately; thus, the response times corresponding to information storage requirements will be greatly minimized without a significant increase in cost over current serial processing machines. In order to achieve this cost reduction for information storage and retrieval, machines with very much larger processing capacity will be required. The trillion byte memories that will be available, whether laser driven or on magnetic tape, will have extremely high transfer rates and will require very large processors in order to support them.

In order to make these extremely large processors cost effective, it is anticipated that national networks will be utilized. Network technology is reasonably well understood at the present time, and the growth of many such networks is anticipated. Large special purpose data bases will be accessible at reasonable costs from any point in the country. Already there are commercially available nationwide networks for special purpose on-line information storage and retrieval. Examples are the Medline system, which contains information derived from medical periodical literature; and the Science Information Association which utilizes Battelle's BASIS 70 for such data bases as the National Technical Information Service bibliographic reference data of unclassified unlimited technical reports and the Chemical Abstracts Condensates data base of journal references in chemistry and chemical engineering.

Due to the rapidly advancing state of the art in information storage and retrieval, and because of the increased emphasis on developing sophisticated equipment for this application, it is anticipated that information storage and retrieval costs in the next three to five years will decrease dramatically.

SECTION 4

NATURAL LANGUAGE QUESTION-ANSWERING SYSTEMS

In Section 2, reference was made to three levels of sophistication of an information system for use by researchers and practitioners in instructional system design. The highest level of sophistication involved an interactive on-line system in which the computer would serve as a consultant. Essentially, such a system would allow the researcher or practitioner to ask the computer a question in normal English. The computer would search its memory store for pertinent details, synthesize them into an answer, and provide the requester with a narrative answer to the question.

In conducting the feasibility study, the concept of such a system was addressed by exploring the state of the art of Question-Answering (QA) systems and artificial intelligence. A comprehensive literature review was performed. The conclusions derived from the comprehensive literature review and the implications for an interactive on-line prescriptive guidance information system are presented in the remainder of this section.

Thus far, no system has been designed that can "understand" natural language at a significantly high level of sophistication. The basis for failure apparently lies in the lack of an adequate theoretical description of language itself.

Recent attempts have been made to formalize linguistic descriptions and to develop theoretical frameworks within which languages can be described. At the present time, however, there exists a great deal of confusion in the field regarding the status of theories which have been proposed. In particular, the primary descriptive theory of language since 1957, transformational-generative grammar, has been subjected to criticism. The goals of transformational grammar were expressed by Chomsky² as follows:

The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of grammatical sequences.....

These goals have been attacked for not being broad enough in scope to lead to future possible breakthroughs in linguistic theory. Therefore, a "revisionist" theory of transformational-generative grammar has been de-

2. Chomsky, N. Aspects of the Theory of Syntax, Cambridge, Massachusetts, The MIT Press, 1965.

3
veloped by McCawley. These neo-transformationalists see semantics rather than syntax as the focal point of investigation. According to these investigators, it is unnecessary and undesirable for a grammar to have both a syntactic categorical component and a semantic interpretation component. A single component was proposed which would generate directly semantic representations expressing meaning and containing sufficient information for the operation of transformations. Thus, as a possible move in the right direction, the "generative semanticists" have emphasized the investigation of the nature of human language and de-emphasized the goal of mechanically generating strings with the correct order and the appropriate constituent structure.

Given the actual expressed goals of Chomsky's standard theory, it is not surprising that meaning and semantic representation is the aspect of human language which had been given the least consideration. Even the generative semanticists, who place major emphasis on semantic representations as underlying surface structures, regard meaning only from the point of view of descriptive linguistics. Winograd⁴ states that meaning must be seen, not as a function of words and sentences (e.g., of phonological or grammatical entities), but rather of the intention of the speaker and the ability of the hearer to reconstruct this intended meaning. And if one is to discuss meaning and semantic representation with cognitive import, it must be discussed in the context of a theory which attempts to explain human thought processes. Just as one cannot describe syntax in the abstract and hope to achieve a model of how syntactic structures are stored in the mind, one cannot isolate a "linguistic" meaning for sentences which can be described apart from a concern with cognitive structures and expect this linguistic meaning to reflect the actual representation of meaning in the mind.

According to Winograd, the kind of semantic theory that is needed is one which incorporates a theory of human cognition. In fact, any meaningful cognitive theory of language must be dependent on a general theory of human cognition. If one maintains a concern for semantics in conjunction with a concern for the cognitive processes which originate thought, then one can see that a theory of semantics must be separate from, but related to, a theory of grammar.

3. McCawley, E. V. The Place of Language in Information Processing, Alexandria, Virginia: Defense Documentation Center, May 1972.

4. Winograd, T. "Understanding Natural Language", Cognitive Psychology, Volume 3 No. 1, January 1972.

Thus, in spite of the claims which Chomsky has made about his theory seeking to give it cognitive status, neither he nor the theory of generative semantics make hypotheses concerning the structure of linguistic knowledge in the mind. Due to the lack of such a cognition-based theory to serve as a foundation, significant progress is not likely to occur in the near future in developing an artificial natural language conversational man-machine dialog question-answering system.

Considerations of a less theoretical nature are significant for the selection of an IS&R system which can serve nearer-term needs of AFHRL. Storing a piece of information so that it can be retrieved is the basic language problem in information processing today. Advantages and disadvantages of the various approaches available are considered below.

At one extreme, index terms can be chosen from a relatively small list of descriptive words. This allows the searcher to query the file readily, and its simplicity is appealing. A problem arises as the collection increases in size, however, since the amount of information cited under any one term can grow to unmanageable proportions. In addition, since the type of vocabulary involved is of a "static nature", constant revisions of the authorized list are required due to the "dynamic nature" of natural human language. It is known that language is in constant flux: terms are constantly introduced and discarded. This is especially true of scientific and technical literature where the occurrences of newly-coined terms introduced by definitions, or the change of existing terms by means of explication, is a readily observable phenomenon.

The basic drawback to the fixed vocabulary indexing system described above is that converting from one authorized vocabulary to another usually involves a great expenditure of time and effort. Large computer runs are necessary as well as much manual indexing. Furthermore, since a complete conversion from one vocabulary to another is usually impossible, a searcher must know old as well as new terms in order to locate all documents in which he is interested.

A second approach to information retrieval, at the opposite extreme, relies on text-word indexing from actual text of documents or selected portions thereof. This provides great flexibility in meeting new demands of technology but creates for the searcher the task of exhausting alternative means of expressing concepts as they may have occurred in the author's natural language.

Most organizations employ a system that lies somewhere between the two extremes cited above. One problem is that many reference tools necessary to search collections at present are not normally available to the user (such tools as thesauri, identifier lists, open-ended term lists, frequency counts, etc.).

The implementation of on-line search techniques has prompted interest in the use of natural language for information retrieval. Today, many people making queries via on-line systems are not information processing specialists. These users, therefore, have no interest in learning a structured artificial language to query a data base. Thus, widespread use of computers by non-computer users makes the idea of using natural language for information storage and retrieval very appealing.

The review of the literature on natural language question-answering systems in conjunction with our own findings of user needs and available IS&R systems has confirmed our conviction that a text-processing automatic indexing system indeed represents the best choice for an information storage and retrieval system to meet the needs of AFHRL personnel. Although a true artificially-intelligent computer-based interactive information system does not seem feasible in the near term, pending further advances in linguistic theory, some features of a prescriptive guidance system as envisioned by AFHRL are already possible with the computer and IS&R technology available today.

SECTION 5

ESTABLISHMENT AND IMPLEMENTATION OF A MODEL DATA BASE OF COMPREHENSIVE ABSTRACTS

5.1 SYSTEM DESIGN

The Avionics Central system as operated by the Avionics Laboratory is the recommended system. In order to validate the feasibility of this system, several hundred comprehensive abstracts prepared under a separate contract were selected to serve as a data base. Arrangements were made with the Avionics Laboratory to create a model data base of these abstracts. The University of Dayton, in conjunction with the Air Force Human Resources Laboratory, established a system design for the data base. The design of the system was based on the results of the users' needs study described in Section 2.

The text processing automatic indexing feature of the software was used to prepare the inverted files needed for storage and retrieval. Certain features, such as the establishment of a stop word list and a synonym equivalence table, were not used for the model data base. Because of the full-text indexing capability of the software, it was necessary only to provide textual data in the appropriate format as input for the creation of the model data base.

The design of the model literature-based information system for comprehensive abstracts consisted of defining fields or segments for the records to be contained in the data base and the data to be incorporated therein. The design was established in conjunction with the software so that retrieval can be effected by various combinations of segments and various segments can be displayed. This is particularly important in a user-oriented interactive system so that the user can scan selected portions of the retrieved records to establish their pertinence and so that search modifications can readily be made to improve the precision on retrieval.

Avionics Central has established a general purpose set of segments for accommodating Government report literature. This set of segments is basically derived from DD Form 1473 (title page) which must be supplied with all Department of Defense reports. The segments defined and implemented for the test system were fitted into the basic DD Form 1473 framework.

-
5. Schumacher, S. P. Development of a Technical Data File on the Design and Use of Instructional Systems, USAF Human Resources Laboratory, Technical Report AFHRL-TR-73-41, December, 1973.

TABLE 1

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
1	ACNUM	ACCESS NUMBER	Unique identification number of record (abstract) in the data base	00001031
2	UPDATE	LAST DATE OF UPDATE	Date on which the file was last updated, i. e., new records were entered into the file. This field is searchable with arithmetic logic	73/03/16
4	ORIGACT	ORIGINATING ACTIVITY	Organization responsible for performing the work or publication (journal title)	A V Communication Review
8	CLASS	REPORT SECURITY CLASSIFICATION	Government security classification category	Unclassified
12	DECLASS	REPORT DECLASSIFICATION CODE		
16	TITLE	REPORT TITLE	Title of the document	Pictorial Embellishments in a Tape-Slide Instructional Program
20	DNOTES	DESCRIPTIVE NOTES	Notes supplied by the report author, usually describing the nature of the report, e. g., quarterly progress report	
24	AUTHOR	REPORT AUTHOR	Personal author(s) of the document	Popham, W J
28	DATE	REPORT DATE	Date document was issued	69/00/00
32	PAGES	NUMBER OF PAGES	Number of pages in the document	0008
36	REFERENCES	NUMBER OF REFERENCES	Number of references cited in the document	0001
40	CONTR/PROJ	CONTRACT, PROJECT, TASK, WORK UNIT	Identification numbers as applicable; primarily for Government reports	

TABLE 1 (Cont'd)

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
44	REPORTNR	ORIGINATOR REPORT NUMBER	Identification number assigned by the originating organization	
48	OTHERNR	OTHER REPORT NUMBER	Other identification numbers, such as AD number, sponsors report number, etc.	
52	DISTRIBUTE	DISTRIBUTION STATEMENT	Statement indicating any limitations regarding distribution of the document	
56	SNOTES	SUPPLEMENTARY NOTES	Any explanatory notes supplied by the author regarding the document	
60	SPONSOR	SPONSORING ACTIVITY	Organization sponsoring the work reported	
61	TYPPUBL	TYPE OF PUBLICATION	Category of publication medium, e.g., Government report, book, journal article, slide set, movie, instructional course	Journal Article
62	CITN	CITATION	Complete bibliographic citation	Popham, W. J. Pictorial Embellishments in a Tape-Slide Instructional Program, "AV Communication Review", 1969, "17", 28-35.
63	ABSTRTYP	ABSTRACT TYPE AND DESCRIPTION	Specific information provided on the summary and evaluation form accompanying AFHRL comprehensive abstracts indicating Type 1 (general descriptive) or Type 2 (experimental), ISD Topic, and category	Type 2 Research Study (Manipulate Var.); ISD Topic
64	ABSTRACT	ABSTRACT		This study investigated the effect on learning of pictorial or cartoon embellishments in a tape slide instructional program. Previous research indicated

TABLE 1 (Cont'd)
File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
64 (Cont'd)	ABSTRACT (Cont'd)	ABSTRACT (Cont'd)		that such embellishments had no effect on learning but increased learner enjoyment of the program. The implied null hypotheses in this study were: (1) There is no difference in learning between SS who are instructed with materials containing cartoon embellishments and SS who are instructed with materials without cartoon embellishments; and (2) there would be no difference in the enjoyment or liking of the two methods. Method: There were over 100 SS from five colleges or universities who were students pursuing administration credentials. The independent variable was the type of instructional materials: (1) Tape-slide instruction program with cartoon embellishment; (2) tape slide instruction program without cartoon embellishment; and (3) a programmed text, a written version of the tape slide program. Instruction was given in the regular classroom session. In each class the procedure was identical. The SS were told they were participating in an investigation of program presentation mode. SS were randomly assigned to one of the three conditions and all subjects received their corresponding instruction programs. Upon completion of the programs, SS were administered a posttest. The first part of the test was a set of multiple choice questions to measure cognitive achievement. The second part was a four-item questionnaire on which each S rated his program on learning, enjoyment, interest,

TABLE 1 (Cont'd)

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
64 (Cont'd)	ABSTRACT (Cont'd)	ABSTRACT (Cont'd)		and willingness to complete similar programs. It should be noted that this experimental design was used for three different instructional topics: improved educational programs, educational criterion measures, and classifying educational research studies. Each topic used a different set of three experimental programs and posttests. Data was analyzed for each topic individually. Results: An ANOVA was used to test for differences in achievement test performance in each of the three topics. One significant difference was found (P .LT. .01). This occurred in the "Classifying Educational Research Studies" program where the unembellished tape-slide program group performed markedly better than the groups instructed by the other two versions. Twelve chi-square analyses on data from the second half of the posttest yielded one significant difference, namely on the enjoyment rating of "Improved Educational Programs." Mean affective ratings of the programs ranged around three on the 5-point scale. Conclusions: **There was no consistent superiority, either on the cognitive or affective measures favoring one of the three treatment methods. ** Although the unembellished version resulted in superior test results in one instance this superiority was not present for the other two programs. Affective results were mixed, with each treatment condition receiving highest ratings by learners on several dimensions. The failure to detect significant differences

TABLE 1 (Cont'd)

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
64 (Cont'd)	ABSTRACT (Cont'd)	ABSTRACT (Cont'd)		among the treatments with respect to achievement test performance in the case of two programs is consistent with findings in an earlier study (Baker and Popham 1965). The exception in the case of one program can probably be attributed to the particular attributes of the program in question. Evaluation--This article was somewhat confusing in that the author did not make clear the reason for including a programmed text in the experimental design. The sampling procedure used in this study was far from random. In addition, the NS for each of the nine categories in the design ranged from 10 to 29. Finally it is doubtful that the procedures used were identical in each case. Because of problems with sampling and control this study is of questionable value to the problem area. The topic is somewhat relevant to ISD. However, the problem of whether to include cartoon embellishments is not particularly important especially because of lack of evidence that such embellishments are helpful. (R=1, P=8, EMW
65	CONCL	CONCLUSIONS	The conclusions of the study are presented in this segment as well as in the body of the abstract so that they can be displayed separately	Conclusions: Conclusions: **There was no consistent superiority, either on the cognitive or affective measures, favoring one of the three treatment measures. ** Although the unembellished version resulted in superior test results in one instance, this superiority was not present for the other two programs. Affective results were mixed, with each treatment

TABLE 1 (Cont'd)

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
65 (Cont'd)	ABSTRACT (Cont'd)	ABSTRACT (Cont'd)		condition receiving highest ratings by learners on several dimensions. The failure to detect significant differences among the treatments with respect to achievement test performance in the case of two programs is consistent with findings in an earlier study (Baker and Popham, 1965). The exception in the case of one program can probably be attributed to the particular attributes of the program in question.
66	EVAL	EVALUATION	The abstractor's evaluation of the document is presented in this segment as well as in the body of the abstract, so it can be displayed separately	Evaluation: Evaluation--This article was somewhat confusing in that the author did not make clear the reason for including a programmed text in the experimental design. The sampling procedure used in this study was far from random. In addition, the NS for each of the nine categories in the design ranged from 10 to 29. Finally, it is doubtful that the procedures used were identical in each case. Because of problems with sampling and control, this study is of questionable value to the problem area. The topic is somewhat relevant to ISD. However, the problem of whether to include cartoon embellishments is not particularly important, especially because of lack of evidence that such embellishments are helpful. (R=1, P=8, EMW)

TABLE 1 (Cont'd)

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
68	KEYWORDS	KEYWORDS	Keywords are assigned to represent the document content by the abstractor; the keywords are based on a controlled listing of allowable terms.	Keywords: Aids, Instructional; Attitude; Auto-Instruction; Criterion Test; Evaluation, Course; Learning, Facilitation; Materials, Preparation; Media, Effects of; Media Instructional; Post-Test; Presentation Methods; Printed Material; Programmed Instruction (PI); Recall; Research, Media; Slide Projection; Sound/Slide; Visual Materials; Visuals, Use of
69	UNIQWDS	UNIQUE KEYWORDS	Unique keywords are assigned by the abstractor to supplement the authorized controlled vocabulary keywords.	Unique Keywords: Cartoon Embellishment; Southwest Regional Laboratory for Educational Research and Development (SWRL)
72	SYMBOL	LOCATION SYMBOL	Symbol to show the physical location of the parent document or hardcopy abstract	AFHRL/AST
76	SAFE/CAB	LOCATION FILE	Code to show specific file cabinet or safe for classified material	
80	COMMENTS	COMMENTS	Comments to be added by user as appropriate	Type: Applied; Field Study; Method: Quasi-Experimental
100	RSCHMETH	RESEARCH METHOD	Statement of research method employed in experiment. This information is derived from the summary and evaluation form prepared in conjunction with the comprehensive abstract	
104	APPAR/MEDIA	APPARATUS/MEDIA USED	Describes apparatus and/or media used in the study	Tape Recorder; Filmstrip (Silent); Sound/Slide Projector; Programmed Text

TABLE 1 (Cont'd)

File Design for AFHRL Data Base of Comprehensive Abstracts

Segment Number	Segment Acronym	Segment Heading	Definition	Example
108	SUBJPOPN	SUBJECT POPULATION	Describes the characteristics of the subject population	100 Normal Mixed College Students Aged 20-30
112	INDVAR	INDEPENDENT VARIABLE	Describes the independent variables or stimuli utilized in the experiment	Type of Instructional Materials
116	DEPVAR	DEPENDENT VARIABLE	Describes the dependent variables or responses studied	Effect on Learning
120	MEAS/STAT	MEASUREMENT/ STATISTICAL METHODS	Describes the measurement techniques and/or statistical methods used in the study	ANOVA; Chi-Square Analysis
124	ADDNLRSCH	ADDITIONAL RESEARCH SUGGESTIONS	Describes the abstractor's suggestions for additional research based on the document	
128	CROSSREF	CROSS REFERENCES	Provides cross references to other items, either in the data base at hand or to items in other data bases	
132	QUALINDEX	OVERALL QUALITY INDEX	Abstractor's assessment of the quality of the document derived from the summary and evaluation form accompanying the comprehensive index. Values are 1, 2, or 3.	

with additional specialized segments to meet the needs of AFHRL. The specific segments, segment definitions and typical specific data are presented in Table 1.

5.2 RETRIEVAL FROM THE MODEL DATA BASE

Retrieval from the data base is accomplished using the Avionics Central system.* Specific user instructions are provided as Appendix B to this report. There are two basic operations in retrieving information: the specification to the system of the retrieval requirements and the display to the user of the records or specified portions of the records. Because of the full-text index/retrieval capability of the system and the file design by segments, the user has a vast range of capabilities available to him. The comprehensive abstracts prepared as input to the model data base are particularly amenable for providing a wide scope of information to the user.

Since the user can display such items as the conclusions, the independent variable, the dependent variable, the subject population characteristics, and other elements, the system has available many desirable features. Certainly the system designed is far superior to a system which would merely generate a bibliography of access numbers, perhaps in conjunction with titles and authors.

Indeed, the system design is entirely adequate to accommodate the computerized handbook concept described in Section 2. The computerized handbook concept could be implemented by incorporating prescriptions as an additional data base. These prescriptions would represent human-derived extractions and syntheses from textbooks, technical reports, and other source material which could easily be cross referenced. Retrieval of prescriptions could be effected by searching the appropriate segment for that data base. Corresponding descriptive keywords could also be incorporated to aid in retrieving the appropriate prescription(s).

5.2.1 Computer-User-Dialog

The first step in searching the data base is to "sign on". The signing on is accomplished by dialing the computer's telephone number. The sign-on dialog is indicated by showing computer messages in all capital letters and user responses in all lower-case letters. Control is returned to the computer (transmit) with a teletype terminal by depressing the control key and the letter 's' simultaneously. This is indicated by (cntrl s). As

* See Appendix A for a description of the Mead Data Central software used by Avionics Central.

soon as the signal is established the computer message is: YOU ARE NOW IN COMMUNICATION WITH AVIONICS CENTRAL (000). PLEASE ENTER 10 CHARACTER IDENTIFICATION.

(002)

REPLY/xxxxxxxxxx(10 character i.d. number)

ENTER FILE, MESSAGE OPTION. (040)

REPLY/afhrl,lkwic (cntrl s)

ENTER REQUEST. (048)

REPLY/\$any equ 'self*pacing' or 'pacing, self-' and instruction (cntrl s)

11 ANSWERS, HOW DO YOU WANT TO PROCESS THEM:

NO, PRINT, MODIFY, OUTPUT, OR ENTER OUTPUT AND DEVICE. (151)

REPLY/99,c (cntrl s)

ENTER SEGMENTS TO BE DISPLAYED

REPLY/1, 16, 24, 28, 62, 65, 66, 68, 69 (cntrl s)

DO YOU WANT THE ENTRIES SEQUENCED? YES OR NO. (211)

REPLY/no (cntrl s)

SET PAPER (IF NECESSARY), PRESS SPACE BAR TWICE AND TRANSMIT. (225)

space space (cntrl s)

Following this dialog, the computer prints at the console those segments specified by the user for all the qualifying documents. When the printout is completed the computer responds:

END OF OUTPUT FOR THIS QUERY. (227)

REPLY/ (cntrl s)

AFHRL IS THE CURRENT FILE, MESSAGE OPTION. ANSWER YES TO CONTINUE, OR ENTER FILE AND MESSAGE OPTION. (047)

REPLY/ yes (cntrl s)

ENTER REQUEST. (048)

At this point, the user has executed a complete search cycle and is now ready to proceed to the next query. When he is finished, it is only necessary to hang up the phone, and the user is automatically logged out. Further details of search and display procedures are given in the following paragraphs. A more complete description of user procedures is given in Appendix B.

5.2.2 Search Procedures

The user initiates a search of the data base by using \$ sign as a "look up" operator. The segment is then indicated by acronym or segment number. An operator is then specified which tells the system how to search. The argument represents the particular value or text components of the segment to be matched in the search process. Logical AND and OR expressions are used to connect individual search arguments. The expression "\$ANY" instructs the system to look in all segments for a match. Examples of search specifications are shown below.

Example 1. Give me documents concerned with the subject
ADJUNCT PROGRAMMING

Search specification:

\$any equ 'adjunct programming'

The search specification tells the system to search the data base in all segments for documents containing the search phrase ADJUNCT PROGRAMMING. Note that search phrases are enclosed in single quote marks.

Example 2. Give me documents on PROGRAMMED INSTRUCTION
authored by L. BRIGGS

\$any equ 'programmed instruction' and \$author equ 'briggs, l'

Example 3. Give me documents in which a TACHISTOSCOPE
was used to study AIRCRAFT RECOGNITION

\$any equ (aircraft or airplane) and recogni****

and \$appar/medi equ tachistoscope

Note that the asterisks permit a word stem search to include any word whose first seven characters are RECOGNI; thus documents containing recognize, recognition, and recognized, would all qualify. This search is broader in nature than if the phrase 'AIRCRAFT RECOGNITION' had been specified.

At the completion of the computer search, the user is notified of the number of documents that satisfy his request. If the number appears reasonable, he can continue with the display operation. However, the number of documents may be rather large. In this case, the user can modify the original request to narrow its scope. For example, by searching \$ANY EQU 'PROGRAMMED INSTRUCTION', 130 documents qualify. There are several means of narrowing the request. One means is to add logical AND words or phrases. Another method would be to require 'PROGRAMMED INSTRUCTION' to be in the title. Still another method would be to require the date to be 1972 or 1973. In the particular instance cited, the search was modified as follows:

REPLY/ \$any equ 'programmed instruction' (cntrl s)

130 ANSWERS, HOW DO YOU WANT TO PROCESS THEM:

NO, PRINT, MODIFY, OUTPUT, OR ENTER OUTPUT AND DEVICE. (151)

REPLY/modify (cntrl s)

ADD 002 MODIFICATION. (049)

REPLY/ and \$title equ 'programmed instruction' (cntrl s)

19 ANSWERS, HOW DO YOU WANT TO PROCESS THEM:

NO, PRINT, MODIFY, OUTPUT, OR ENTER OUTPUT AND DEVICE

REPLY/ 00,p (cntrl s)

This additional restriction resulted in only 19 documents, a reasonable number to examine. It was assumed that those documents actually con-

taining the word phrase 'programmed instruction' in the title should be more relevant than those documents merely mentioning programmed instruction in the abstract or in some other segment. The reply 00,p (print all segments at the printer) directed that the complete records of the qualifying documents be printed off-line and forwarded to the user; the address of the user is maintained by Avionics Central.

Another possibility is that no documents qualify in response to the request. In Example 3 above, on tachistoscopes used for aircraft recognition there were no documents which qualified. The original request was modified as follows:

NO ANSWER FILLS REQUEST. (150)

REPLY/cntrl s

ENTER FILE, MESSAGE OPTION. (040)

REPLY/afhrl,lkwic (cntrl s)

ENTER REQUEST. (048)

REPLY/\$appar/medi equ tachistoscope (cntrl s)

YOUR REQUEST IS BEING PROCESSED (109)

2 ANSWERS, HOW DO YOU WANT TO PROCESS THEM?

NO, PRINT, MODIFY, OUTPUT, OR ENTER OUTPUT AND DEVICE. (151)

REPLY 00,c (cntrl s)

SET PAPER (IF NECESSARY), PRESS SPACE BAR TWICE AND TRANSMIT.
(225)

REPLY space space (cntrl s)

The result of this output was to show the complete records of both documents in which a tachistoscope was used as the apparatus/medium. Even though the application of the tachistoscope was other than aircraft recognition, valuable information on the use of tachistoscopes was obtained. Another search was run on aircraft recognition; this search resulted in one document. By combining the information contained in the results from these two searches and by referring to the reference lists of the retrieved documents, the researcher would have a good start in finding relevant material on the appropriateness and usefulness of tachistoscopes for train-

ing personnel in aircraft recognition. It should also be noted, of course, that the model data base searched consists only of about 500 abstracts, which can account for a low number of retrievals.

Keywords can be very useful for searching. Keywords used in manual indexing are derived from a controlled vocabulary with a fixed format, and therefore represent uniformity both in format and with regard to the semantic representations of the keywords.

5.2.3 Display

One of the chief advantages of the Avionics Central model data base system is the wide number of display options available to the user. Since the Avionics Central system is interactive, the display features can be used to great advantage in screening the retrieved materials for relevance. By successive search modifications, the user can sharpen the precision of results until he achieves exactly that subset of documents from the data base which he wants. This technique of retrieval display, modification, display, etc. is called the interactive iterative (I^2) retrieval technique.

For the first screening, only access numbers and titles may be displayed. From this screening, obviously nonrelevant documents often can be spotted. Also, recurring factors causing nonrelevant retrievals may suggest terms or phrases which can be negated in a search modification for nonteaching purposes, e.g., student records, may formulate a search with \$ANY EQU COMPUTERS AND EDUCATION OR TRAINING AND \$ANY NEQ CAI OR COMPUTER-ASSISTED INSTRUCTION. The expression "neq" means "not equal to". As the I^2 technique derives the set of documents which looks promising, the researcher can request more segments to be displayed for more detailed screening or, indeed, to provide the information he wants directly. If desired, the complete records can be ordered off-line and sent to the user. The user is cautioned that complete records or lengthy segments such as the abstract do require considerable time to be printed or displayed on-line, especially if there are very many documents. With CRT display terminals, the "paging" feature permits the user to browse through records very conveniently. The paging feature refers to the capability of displaying records at various intervals, for example, every tenth document record.

In addition to displaying whole records or segments of records, the user can enter a display command for only those segments in which the search words were found (HITS), segments in which the words were found plus additional segments (HIT-N-PIC), words in context around the search words in whatever segment the search word appears (KWIC-IT), and words in context around the search words in whatever segment the words appear

plus additional segments (KWIC-N-PIC). Further explanation is provided in Appendix B. An example is given below for the phrase "teaching machine".

AFHRL LIBRARY LOCATION

AC 20283 2 (1 of 2 ans.)

ACCESSION NUMBER:

00000283

REPORT TITLE:

HIERARCHICAL PREVIEW VS PROBLEM ORIENTED REVIEW IN LEARNING AN IMAGINARY SCIENCE

REPORT AUTHOR:

MERRILL, M D; STOLUROW, L M

CITATION:

MERRILL, M.D., & STOLUROW, L.M. HIERARCHICAL PREVIEW VS PROBLEM ORIENTED REVIEW IN LEARNING AN IMAGINARY SCIENCE. "AMERICAN EDUCATIONAL RESEARCH JOURNAL," NOVEMBER 1966, "3"(4), 251-261.

ABSTRACT:

... THE MATERIALS WERE PRESENTED BY MEANS OF SOCRATES, A COMPUTER-BASED TEACHING MACHINE USING AUTOTUTOR TEACHING MACHINES AS INTERFACE UNITS (STOLUROW & DAVIS, 1965; MERRILL, 1964). SIX HUNDRED ...

KEYWORDS:

FRAME; LEARNING; PROBLEM SOLVING; RESEARCH; RETENTION; REVIEW; TEACHING MACHINE; TESTING

APPARATUS/MEDIA USED:

CAI/CMI; PROGRAMMED TEXT; TEACHING MACHINES; SELF-INSTRUCTION

SUBJECT POPULATION:

675 NORMAL MALE COLLEGE FRESHMEN

5.3 INDEXING, UPDATING, MAINTENANCE

Avionics Central software performs full text automatic indexing resulting in both inverted files and serial chronological files. Thus, the indexing function is performed directly from text in machine-readable form. If desired, keywords can be assigned manually from a controlled-vocabulary

list. This concept was described under the paragraph above under searching. Keywords are handled by the system just as any other textual data, such as abstract words.

Updating with Avionics Central is a batch mode operation and a time-consuming one. In a computer system such as Mead Data Central there is a trade-off between rapid updating and rapid response on retrieval. Avionics Central has elected to bias the system heavily in favor of rapid response. For a system such as the model data base and for the majority of IS&R systems, this situation is ideal, since updating is performed rather infrequently and rarely are existing records edited.

Updating is accomplished by supplying additional increments of abstract record data to Avionics Central. The mechanics of updating are accommodated automatically by the software. In essence, the inverted files are regenerated, and the serial file is added onto. A very convenient feature is the ability to edit previously existing records to correct mistakes, to insert additional data, or to update with new validated data which supercedes that data which previously existed in the record.

The significant aspect of the updating process is that computer-readable tape in the correct format must be supplied. In the creation of the model data base, MT/ST tapes were converted to machine-readable form, and supplementary data available only in hardcopy were keypunched and merged. The format to be provided to Avionics Central is 80-column card image records as follows:

Col. 1 - 60 Data/text	Col. 61 Blank	Col. 62 - 69 Acc. No. with Leading zeroes (right justified)
Col. 70 - 74 Segment No. (right justified)	Col. 75 - 77 File No. (003)	Col. 78 - 80 Sequence No. (Line no. within the segment for multiline items; right justified)

In initially processing data into the model data base, inconsistencies in input data were present which created many problems in applying corrections. Some guidelines for preparing input data in MT/ST form which should help in future updating are as follows:

1. The MT/ST codes should be applied to permit easy separation into segments as desired.

2. Fields or segments should be separated by clearly identifiable delimiters (something other than spaces). Special characters such as @ would work.
3. The access number should be carried (entered) with each separate field.
4. The MT/ST tape must correspond to the hardcopy. Erasing hardcopy does not change the tape. The MT/ST tape must be corrected.
5. Avoid using lower case l and alpha O to represent numeric 1 and 0. The numeric values must be used when intended.
6. Within the text of abstracts, use a consistent means of indicating points to be made. Inter- as well as intratext consistency is important. As an example, "The following performance criteria were applied:

(1) -----
(2) -----
(3) -----
etc.

7. The identical format must be used in providing data for the various segments; again inter- as well as intratext consistency is required. An example follows:

correct format - AV Communication Review
incorrect format - A V Communication Review
A-V Communication Review

Maintenance of the data base is a function of the operation of the computer system. For the model data base, maintenance is accomplished by Avionics Central. The files are maintained with random access disk storage of data. File integrity is ensured by locking of files through the identification number.

5.4 COSTS

The costs of maintaining, updating and obtaining access to the data base are not currently available. It may be reasonably assumed, however, that the costs of operating a full-scale data base or data bases through the Avionics Laboratory would be significantly less than with a commercial arrangement, either with Mead Data Central or with some other vendor.

Assuming a data base of comprehensive abstracts similar to the model data base, estimated annual costs are broken down assuming the commercial rates for Mead Data Central which are known. To fit within the three to five year time period it is assumed that about 20,000 abstracts may be available for the data base.

Software/hardware leasing including user-system interactions	\$1500/mo. x 12 mos.	\$18,000
Storage of data on discs	\$10/10 ⁶ characters/ mo. x 100 · 10 ⁶ char. x 12 mos.	\$12,000
Conversion of MT/ST tapes	\$1000/update x 4 updates/yr.	\$ 4,000
Preparation of computer- readable update tapes and updating of data base through Avionics Central, assuming quarterly up- dates		\$ 2,000
Total annual cost		\$36,000

It should be noted that the costs of the information system do not include the costs of preparing the comprehensive abstracts. It should also be pointed out that the comprehensive abstracts represent a data base of document surrogates from which one actually can retrieve facts and principles as well as evaluative and reference data. As such, the data base is unique, since other IS&R systems do not provide the depth of analysis and comprehensive-ness of coverage in the field of training and educational research as well as actual instructional system design.

Considering the value of such a data base, not only to the Air Force but to many potential users in the training community, it seems reasonable to consider amortizing the cost of maintaining and updating the system. This could be accomplished by making it available not only to other Air Force and military organizations, but also to the aerospace industry, psychologists, universities, the Department of Health Education and Welfare, and other organizations. By charging for services, cost sharing would be achieved, just as Avionics Central is cost sharing Mead Data Central software and hardware among many users.

REFERENCES

1. Scheffler, F.L., A.T. Dodson, T.R. Featheringham, and J.P. Webb, A Survey of the Technical Information Environment of the Air Force Flight Dynamics Laboratory, UDRI-T-69-29, University of Dayton Research Institute, Dayton, Ohio, October, 1969.
2. Chomsky, N. Aspects of the Theory of Syntax, Cambridge, Massachusetts: The MIT Press, 1965.
3. McCauley, E.V. The Place of Language in Information Processing, Alexandria, Virginia: Defense Documentation Center, May 1972.
4. Winograd, T. "Understanding Natural Language", Cognitive Psychology, Volume 3 No. 2, January 1972.
5. Schumacher, S.P. Development of a Technical Data File on the Design and Use of Instructional Systems, USAF Human Resources Laboratory, Technical Report AFHRL-TR-73-41, December 1973.

APPENDIX A

DESCRIPTION AND ANALYSIS OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS INVESTIGATED *

<u>IS&R System</u>	<u>Page</u>
Document Processing System	52
Mead Data Central	56
Aerospace Materials Information Center	61
BASIS-70	65
Remote Information Query System (RIQS)	69
DIALOG	74
ORBIT II	77
NASIS	80
Query/Update	85
STAIRS	88

* The evaluation of various IS&R systems and recommendations were made solely by the contractor. The U. S. Air Force does not assume any responsibility for the contractor's comments on the various systems described and evaluated, nor does the Air Force necessarily endorse the contractor's views.

DOCUMENT PROCESSING SYSTEM

1.0 DESCRIPTION

1.1 INTRODUCTION

The IBM Document Processing System (DPS) operates under control of Operating System/360. The system is designed to process narrative and bibliographic data into interrelated data sets. Searching is done in an on-line interactive mode, and positional or word phrase logic as well as normal Boolean logic is available.

1.2 DATA BASE CREATION

DPS accepts data in natural language form. Each entry consists of a sequential accession number, followed by bibliographic data (optional-up to 249 elements), followed by the text. The maximum numbers of words per document is over 65,000. Data can be input using any medium (tape, cards, etc.) available to OS/360.

Input data is matched against a Common Word file (up to and including 255 words with a maximum word length of 8 characters) to delete semantically insignificant words, e.g., THE, OF, BUT, IT. The input is then matched against the Dictionary File, and nonmatches with the Dictionary are printed out as non-found words. These terms can be reviewed by lexicographers and added to the Dictionary File as desired. The data is stored both in the inverted Vocabulary file and in the Master File. The Vocabulary File contains numeric equivalents of Dictionary terms along with the associated access numbers. The Master File contains the bibliographic reference and positional data which indicate the relative positions of words in the input text.

1.3 USER INTERACTION

Searching of DPS files is performed in an on-line interactive mode. The user indicates which data base he wants, if more than one data base is available. If desired, the search can be limited to a specific range of DPS accession numbers, such as the most recent update material.

DPS searching is accomplished with a "search mode" and a "qualification mode". In the search mode, any keywords in the Dictionary File can be combined using Boolean (AND, OR, NOT) logic. If a more restrictive search is desired, positional operators may be used to indicate that the requested terms must co-occur juxtapositionally as a word phrase, within a specified number of words, in the same sentence, or in the same paragraph. Multiple levels of logic can be handled by labelling each line of the search, and then combining the labels

with Boolean and/or positional logic. For example:

```
$1 fiber, fibre
$2 glass
$3 $2 & $1 (+1)
$4 polymer & composite (sen)
$5 $3 & $4
```

This search requires that the phrase "glass fiber" or "glass fibre" occur in the same document as a sentence containing the words, "polymer" and "composite". This search could be made broader by using different positional operators, e.g., (+2) in line \$3. The (+2) operators would retrieve the terms "glass" and "fiber" or "fibre" separated by a maximum of one word in either direction. If a narrower search is desired, (+1) could be used in line \$4.

After the search is executed, the number of retrievals is printed on-line: the user then has the option of the qualification mode. If he wants only the very recent unclassified material he may add these lines:

```
if date ge 71          or          $6 if date ge 71
and classif lt 1       $7 if classif lt 1
                        $8 if $6 and $7
```

By labelling his qualifying statements, the user can use multiple levels of Boolean logic among qualifiers.

Output is available either on-line or off-line. The document titles and all bibliographic elements can be listed on-line. Abstracts are not available on-line because of the time needed for printing. However, complete output (title, bibliographic data, abstract text) can be ordered off-line.

1.4 SPECIAL FEATURES

DPS provides for truncation of terms by entering the notation (\$) after the word root, e.g., produc(\$). The truncation feature can only be used in a logical OR string, but by labelling one can readily incorporate a truncated word stem in a positional logic statement as follows:

```
$1 Manufactur($)  
$2 Facility  
$3 $1 & $2(+1)
```

DPS also provides the option of establishing a Synonym/Equivalent data set. These lists contain related search terms to be searched in addition to the requested Dictionary entry. The system designer must decide what words (if any) are synonymous or equivalent to each Dictionary term, and must specify these relationships in the file.

In the Foreign Technology Division (FTD) application for Centralized Information Reference and Control On-Line (CIRCOL), auxiliary files have been established called the Nonsignificant Word File and the Word Form Conversion File. The Word Form Conversion File permits alternative forms to be converted to the standard Dictionary form and the Nonsignificant Word File extends the function of the Common Word File, thus permitting the rejection of more than 255 different words and permitting nonsignificant words longer than eight characters to be rejected.

The Limit Range (LRANGE) feature incorporated in CIRCOL permits one to specify the portion of the file to be searched. Unless the LRANGE is specified, the entire file is automatically searched. By restricting the search to a portion of the file, the turnaround time in the interactive mode is improved considerably. The feature is particularly useful in "negotiating" a search request. For example, if a search strategy of 100,000 documents results in 200 documents and the total file consists of more than 800,000 documents, a search of the entire file will retrieve about 1500 to 1600 documents. Therefore the search strategy should be made more restrictive before searching the entire file.

2.0 HARDWARE

DPS is operating on the IBM 360/65 at FTD. It uses OS and supports teletype speed devices. Response time is rather slow, largely due to the size of the data base.

3.0 COSTS

The costs for DPS would be dependent on arrangements that could be made for installing and operating software. The FTD system is currently unavailable for other data bases outside of FTD.

4.0 CONCLUSIONS

Advantages:

1. The system is an interactive natural language-based system.
2. The system is well capable of accommodating the data base envisioned for the AFHRL applications.
3. The cost would probably be fairly low.

Disadvantages:

1. The software program is fairly old; it was written for second generation hardware and does not operate efficiently on third generation equipment.

2. The turnaround time is slow.
3. Output options are limited; customized output formats must be designed.
4. The Dictionary File, NSW, CW, and WFC Files must be continually maintained by human lexicographic decision-making.
5. Only batch mode updating is possible.
6. Previous records are not editable.

In view of the above, DPS does not warrant further consideration.

MEAD DATA CENTRAL

1.0 DESCRIPTION

1.1 INTRODUCTION

Mead Data Central is a multi-purpose information system developed to allow maximum flexibility in application utilization and file and/or data base design, as well as hardware utilization in an on-line interactive mode. It is not limited to text retrieval or document handling, but is capable of handling all types of data.

1.2 DATA BASE CREATION

The major flexibility of Data Central is that it allows the user to view the data he is processing in his own logical fashion without regard to the physical manner in which it is stored, thus permitting maximum efficiency in the utilization of storage techniques and other technology. This cross correlation between physical and logical data is handled by an internal file within Data Central called the Data Base Definition file. The logical view for the most part can be changed at the option of the user without necessitating the physical restructuring of the file; the physical form of the file may be restructured to take advantage of newer technology and newer equipment without the logical form of the file being affected.

The system provides for multiple level data structures. The internal file structure of Data Central is a completely controlled index sequential structure wherein the data content of the user-defined data base is stored in both a sequential (serial-indexed) and inverted (word-indexed) mode.

Batch use is available for all the non-interactive functions such as data base creation, data base definition, data base updating, and is also available for large volume retrieval displays resulting from interrogation.

Data input is accomplished via any machine-readable medium including tape, cards, Optical Character Reading (OCR), cartridge, etc. Data storage is primarily a direct access mode (disk or data cell) with tape providing backup.

The field definition of data items or segments is performed via the assignment of an alpha-numeric string ($\leq$ 40 characters in length) as the identifier (acronym) and the assignment of a unique "segment number".

1.3 USER INTERACTION

On-line use is available for interactive interrogation and display of information. A user accesses the on-line system through an "access key" which determines those data bases and file of the accessible data bases which he may enter. The on-line interrogation is done with free form (quasi-English) with high level operators [equals (=); less than (<); between (/); greater than (>)] and Boolean connectors (AND, OR, NOT).

In conducting a search there are two basic comparative conditions available: logical conditions and arithmetic conditions. Both require the specification of segment name, operator and argument. The argument is the actual data item. For example a logical condition : \$AUTHOR EQU MILLER means that the word "MILLER" (argument) is to be found in the segment (field) known as "AUTHOR". The operator may be indicated either by the = sign or EQU.

For example an arithmetic condition:

\$LENGTH > 1.0 YD, 1 FT, 8.1 IN

means that the arithmetic value of "1.0 YD, 1 FT, 8.1 IN" (or "56.1 IN" or 4 FT, 8.1 IN") must be exceeded in the segment (field) known as "LENGTH".

The use of Boolean connectors with the system allows for two levels of conjunctivity and one level of disjunctivity. They are defined as follows:

A and B or C means A & (B or C)

A & B or C means (A & B) or C

The use of more than one operator to connect arguments for the same segment name is allowed. For example:

\$DATE BTN JUN 70 AND JAN 72

Data Central is an interactive system, and by using the command MODIFY the user may at any point in the search/retrieval process return to the search mode. By doing so, he is not required to restate his already defined and pre-stored selection criteria.

Each entry in the Data Central data base is defined by a unique Data Central-assigned accession number. If this number is known, access to the data directly (bypassing the normal search criteria) can be made. For example:

\$NR {EQU
BTN} X, X, X, ...X (where X stands for the known accession numbers). If the operator is "EQU" then each specified accession number is set for retrieval. If the operator is "BTN" then the numbers are expected in pairs and each pair represents a range of accession numbers to be retrieved.

Output from Data Central is available in several forms. One feature of the output is the ability to sort the output information as desired. The primary report capability of Data Central lies in the ability of Data Central to interface formatting subroutines written in any of the existing procedural languages or in one generalized subroutine. The system has the capability to insert, at appropriate points in data, codes to effect color display on the currently available color CRT devices. The system also allows the user to skip the remaining data in the report for one entry and move immediately to another entry report (paging).

1.4 SPECIAL FEATURES

Data Central has the ability to depluralize the words found. This assures that both singular and plural form of keywords are selected independent of the form which is specified in the selection criteria. Certain words and word forms can be made equivalent for the selection process, e.g., "teach" can be made equivalent to "instruct".

Data Central also allows for the implied complex specification of segment definitions within a search. For example, assume two segments of a file are "author-location" and "institution-location"; assume also a group segment of both "author-location" and "institution-location". A search specified as "Find any entry whose location is Ann Arbor in the state of Michigan", would be specified as:

Location = Ann Arbor and Michigan

It is considered equivalent to:

(author-location - Ann Arbor and Michigan) or
(institution-location = Ann Arbor and Michigan)

Another special feature is the specification of a distance search. In this specification two logical components can be linked with one of four special operator-connectors as follows:

- a) Single quotation marks about a word phrase requires the word phrase to appear in the text searched to effect retrieval. For example: \$ TEXT EQU 'ALUMINUM ALLOY'.
- b) - (Wn) - this is exemplified by:
\$ PROJECT-STATUS EQU BALLISTIC (W6) MISSILE
With this specified search condition, the two logical components must occur Within six words of each other.
- c) (WMn) - This is exemplified by:
\$PROJECT-STATUS EQU BALLISTIC (WM6) MISSILE
In this condition, the first phrase component must appear within the specified number of words in front of (Minus) the position of the second specified phrase component.
- d) - (WPn) - this is exemplified by:
\$PROJECT-STATUS EQU MISSILE (WP6) BALLISTIC
The first phrase component must appear the specified number of words behind the second phrase component.
- e) - (WPnMn) - or (WMnPn) - This is exemplified by:
\$PROJECT-STATUS EQU BALLISTIC (WP2M4) MISSILE
In this condition the length of each directional distance is separately specified (e.g., BALLISTIC must appear Within Plus 2 words or Minus 4 words of MISSILE).

The use of a universal character (*) in the argument allows for variants in spelling, e.g., SM*TH* implies SMITH, SMYTHE, etc. The use of multiple universal characters appended to a root word allows for root/stem expansion, e.g., TAX***** implies TAX, TAXABLE, TAXPAYER, TAXPAYEE, TAXATION, etc.

An on-line tutorial or Computer Aided Instruction (CAI) capability is available to the user for assistance at the user's option.

2.0 HARDWARE

Data Central operates on any IBM 360/370 equipment under OS or DOS and supports the following terminal types: TTY, 1050, 2740, 2741, 2260, CC-30. The recommended minimum hardware configuration to run Data Central is an IBM 360/40 with 128K core and multiprogramming capabilities. Response time has been in the neighborhood of 5 to 30 seconds depending upon the load.

3.0 COSTS

The costs to use Data Central on Mead Corporation's computer system are approximately \$1500-1800 per month. These costs include: user training sessions, data base training, and an accounting package. It is also recommended that an on-site representative be available for six months to a year at a standard charge of \$15-18 per hour.

4.0 CONCLUSION

We find the following advantages with Data Central:

1. There is good documentation, including the definition of terms and examples.
2. There is a competent local supplier.
3. The interactive language is free form (quasi-English) with Boolean connectors.
4. It is a comprehensive system.
5. The system has on-line and batch capabilities.
6. There is a tutorial mode to assist users.
7. The user need not worry about internal storage structures.
8. The system is already operational on the Wright-Patterson Air Force Base Avionics Laboratory computer system.
9. Editing capability for data (text) already in the file is provided.

We find the following disadvantages:

1. It may be an oversophisticated system for the application for which it is being considered.
2. Updating is done only in batch or background mode. (This could be an advantage by precluding record modifications by unauthorized individuals).
3. The cost is somewhat high, although through special arrangements with the Avionics Laboratory it may be possible to reduce costs.

THE AMIC SYSTEM

1.0 DESCRIPTION

1.1 INTRODUCTION

The AMIC system is an inverted file information system which is keyword code oriented. The system operates only in a batch mode, although on-line capabilities could be added through additional programming. The system is readily adaptable to numerous applications by appropriate coding techniques.

1.2 DATA BASE CREATION

The AMIC data base is created by the input of keywords (phrases and single terms), bibliographic parameters, data ranges, or any other descriptive items specified by the system designer. The input is provided in a specialized machine-readable format. The required input format and the data base characteristics are as follows:

- a) A human-assigned seven-digit access number followed by a single-or-two-character alphabetic or alphanumeric code; the code can serve as a subject category, field designator, a link-role designator or it may have some other meaning ascribed by the system designer.
- b) The actual keyword, data range, or bibliographic item itself. The maximum allowable length is 70 characters. The number of items per access number is unlimited.
- c) The maximum number of keywords (inclusive of data range, bibliographic items) is 9,999,999.
- d) The maximum number of access numbers which can be input is 9,999,999.

The input processing consists of four steps: Master Word List match; generic term generation; data inversion; sort and merge with the previously existing file. The access number-data item (or keyword) combination is first matched against a Master Word List (MWL) which checks the validity of the entry and converts valid entries to a numeric code. Nonmatching (hence, nonvalid) entries are printed out and returned to the indexer/editor for resolution.

Access number-code number units are passed through an automatic hierarchical/synonym generating program which automatically creates broader hierarchical classes and equivalent synonyms for the original input keyword, depending on the system designer's specifications for this file.

An inverted file is established for the particular update run. This file is in the same format as the main file and can be searched just as the main file. Since the newly-established inverted file consists only of update material, this file can be searched to provide current awareness or Selective Dissimination of Information (SDI) output for the users; this can be done automatically.

The inverted file of update data is added to the previously existing data base by a sort/merge procedure. Any duplicate entries are automatically eliminated. To conserve search time, the search file can be partitioned into large segments, usually by date, so that older material is "semi-retired" and is searched only on demand. In the AMIC system, five-year increments of the data base are maintained in the "active" search mode.

1.3 USER INTERACTION

The AMIC system is a batch-mode system. Boolean AND, OR, NOT keyword (data item) linkages are provided. The user formulates a search strategy by preparing a listing of keywords with the appropriate Boolean operators. Multilevel Boolean logic is provided. The search strategy is submitted and the results are returned; the documents which qualify according to the search strategy are displayed by access number/one- or two-character code. The actual search strategy is also displayed. A file of abstracts, index cards and documents in access number order should be maintained on-site for easy access and reference to the document collection.

1.4 SPECIAL FEATURES

Complete editing capabilities are provided. One can delete keywords and all associated entries, remove a specific access number from all its keyword postings, and delete a specific access number from a specific keyword.

The search program incorporates a "cut-off" feature which permits the document retrievals to be sorted in an ever-increasing degree of specificity. Since on-line negotiation is not available, this feature permits the user to have the effect of running several searches at once, each at a more precise level. In the event that the most precise level may have 0 hits, the other levels can be examined at the user's discretion. Without this feature, a search either would have to be re-run with successively less precise strategies, or the most general strategy would have to be run first, thus virtually guaranteeing low relevance and a large-scale screening task. Thus the cut-off feature simulates to some extent a capability usually available only with on-line systems.

2.0 HARDWARE

The AMIC system operates on the CDC 6600 utilizing the System Indexed Sequential (SIS) file structure and the CDC 6600 operation system (Scope 3.3). This system is maintained at the Aeronautical Systems Division (ASD) computer facility. The minimum core capacity would consist of about 165K (Octal).

3.0 COSTS

The costs of using the AMIC system would be minimal if used on the ASD Computer Center. Arrangements would have to be made with the Center to operate the system. It is estimated that an initial investment of about \$10,000-25,000 would be required to design the system and to initiate the file structures required; the cost would depend on the complexity of the system. It is assumed that the system design and start-up could be accomplished under contract. System maintenance, updating and searching would cost about \$500 per month.

4.0 CONCLUSION

We find the following advantages with the AMIC system:

1. This system is undoubtedly the lowest cost automated system available.
2. There is a competent local supplier.
3. The system has proved highly effective, even for rather complex strategies.
4. Thesaurus generating software is provided with the system.
5. The system is one of the few available with automatic hierarchical posting capabilities.
6. SDI output can be provided directly with each update.
7. The files are completely editable.
8. The cut-off feature simulates to some degree on-time interaction.

We find the following disadvantages:

1. Only batch mode is currently available.
2. Manual assignment of authorized keywords (indexing) is required.
3. There is a limited number of keyword/data item elements available, although this restriction is not severe (10 million).

4. The searching must be performed by manual selection of keywords from the authorized keyword/data item listings; a subsequent search term linking by Boolean operators for a basically one-time search is then required.
5. Only Boolean logic is available.

BASIS-70

1.0 DESCRIPTION

1.1 INTRODUCTION

The Battelle Automated Search Information System (BASIS 70) is an interactive information storage and retrieval system. It has been operational since July 1970 within several of Battelle's large scientific and technical information analysis centers, and nearly 30 data bases have been loaded into the system. BASIS-70 is completely user oriented and no programming experience is needed to use the system.

1.2 DATA BASE CREATION

The particular data base may consist of items such as journal articles, reports, letters, memos, interviews, data groups, etc. Each item is, in turn, divided into fields or data base elements, such as, accession number; title or subject; author; data; company; text; index terms; actual mechanical and physical properties values; and other data points or data groups. Items may be retrieved by the appropriate use of any search term(s) which, depending on the information file, might consist of keywords, authors, facilities, free-text words, or any other class of terms especially suitable as descriptors, data points or data intervals. The designer of the particular data base must specify allowable index/retrieval terms.

1.3 USER INTERACTION

On-line use via a teletype or similar device is the normal manner in which users interrogate the data base. After dialing into the system, a request is made for a valid USER NAME and valid PASSWORD to ensure the security of the BASIS-70 files. Both the valid USER NAME and valid PASSWORD are unique to each user and must be obtained prior to operating the system. The USER NAME and the PASSWORD are blanked or masked by the system so that they are not visible to others. If the system cannot recognize both codes as valid, the user cannot continue.

After a user has logged on to the system with a valid USER NAME and PASSWORD, further security can be obtained for a particular user group and corresponding files by using a sequence of random numbers, which would follow the acronym of the data base. Only those designated to use the file would have the proper name and number sequence. An on-line orientation to the system is available for the user who desires additional operating instructions. In conducting his search, the user retrieves items pertaining to the desired index terms through the use of logical Boolean statements and the DISPLAY command.

The use of logical statements involves the combination of line members representing the desired search terms with appropriate connectives (AND, OR, NOT).

For example:

1/TITLE MISSILE

2/TYPE SAM

3/RANGE (1000-2000) Note: for this data base; the system finds items within a quantitative range of 1000 to 2000 miles

4/(1 AND 2 AND 3)

At any stage during the query, the system allows the user to retrieve items associated with line numbers, either by single search terms or by logical combinations of search terms. This can be accomplished by means of a DISPLAY command.

For example:

6/DISPLAY 4

1.4 SPECIAL FEATURES

Should an invalid search term be entered, either as a result of misspelling or incorrect form this system will immediately notify the user, and provide an option for display of NEARBY TERMS. In addition to listing the NEARBY TERMS, the system also displays the number of items associated with the terms.

The STEM option provides the user with the capability to PRINT those search terms within the data base which contain the desired stem.

For example:

/ABLAT*

or

/ABLAT*20

If the user terminates the stem with an asterisk, no more than ten terms containing this stem will be displayed. The user has the option of changing this limit from ten by entering the desired limit of terms, n, after the asterisk.

As the user proceeds with his search, he will frequently find that the terms and previous logic statements which he currently wants to co-ordinate for document retrieval are scattered about the printout or are no longer on the display screen when using a CRT. Since the line numbers associated with the term and previous logic statements are required to generate new logic statements,

it is imperative that the user have rapid access to them. The LIST option assures this capability. The format of this command is to type the word LIST followed by ALL or the line numbers desired separated by commas.

For example:

```
/LIST 1, 3, 4  
      or  
/LIST ALL
```

The RESTART command is used to switch to a different information or data base without going through the LOGIN procedure a second time.

For example:

```
/RESTART
```

The result is that all previous statements, index terms, logic combinations and commands are erased and the user may start with Line Number 1 for whichever file is desired.

Current developments that Battelle is reviewing with regard to BASIS-70 are:

- a) On-line updating, purging, and editing of files so that qualified users can conduct their own file maintenance.
- b) Development of an interface which will permit simultaneous interaction with computer files and a microform storage device from a single CRT terminal. Random access to indexes stored on high-speed disks and linked to archival data stored on microform media will significantly reduce the overall costs of operating and maintaining massive information files in an on-line mode.
- c) On-line generation of a display of graphic data.

2.0 HARDWARE

BASIS-70 is operated on Battelle's CDC 6400 computer via the Intercom Timesharing Operating System and supports teletype speed devices or CRT devices. Response time has usually been less than 5 seconds.

3.0 COSTS

The rental costs of BASIS-70 are \$1350/mo. for 4 hours per day service and \$2,000/mo. for all-day service. These prices assume a 25 million character data base.

The system could be installed on the WPAFB CDC-6600 for approximately \$25,000-\$35,000. Costs to load a data base range from \$10,000-\$12,000 and updating the data base costs between \$250-\$300/million characters.

4.0 CONCLUSIONS

We find the following advantages with BASIS 70:

1. There is a near-local competent supplier.
2. The system has on-line and batch capability.
3. Response times are good.

We find the following disadvantages:

1. The cost is high.
2. Truncated stem searching is not available. The truncation only permits display of allowable terms.
3. The system appears to be somewhat cumbersome to operate.

RIQS

1.0 DESCRIPTION

1.1 INTRODUCTION

The Remote Information Query System (RIQS) is a generalized information storage and retrieval system developed at the Vogelback Computing Center, Northwestern University which operates in both batch and on-line processing modes. Partial support for the development of this project came from the Air Force Office of Scientific Research.

The RIQS system is designed to create a data base of like structured records of medium size; it is capable of updating this data base and of searching it for records which satisfy certain conditions. Printed reports can be generated from the records found. RIQS is also capable of performing numerical calculations upon records, printing the results of those calculations, or selecting records depending on the results of these calculations. Alphabetically sorted indexes can be generated from the whole file or from just those records satisfying a particular search. The system is not designed for applications involving large amounts of numerical calculation, although it does include numerical calculation capabilities. It is also not designed to handle files which contain records which are very small for example, consisting of only a few words.

1.2 DATA BASE CREATION

Creating a file involves defining what items of data are in a record, what conditions of acceptance apply to that data, and finally, entering the actual data. The user begins by using the CREATE command followed by the chosen file name, which is followed by the word FILE. Once the file is created, the user may retain it on the disk for reuse at a latter date. Protection of the file is available through the password feature of permanent files.

The RECORD DEFINITION phase which follows the CREATE command defines the items of information that make up the records of the file. For example:

```
RIQS
Create Example Abstract File
( 1) Reference Code
( 2) Publication Name
( 3) Volume Number - Issue Number
```

- (4) Date of Publications
- (5) Beginning Page
- (6) Title of Article
- (7) Author
- (8) Author Professional Affiliation
- (9) City and State of Affiliation
- (10) Abstract of Articles
- (11) Author-supplied keywords
- (12) Classification categories assigned

In the record definition phase, each item of the record is assigned a name and an item number. The item number is used as an abbreviation for the name. After item numbers and names have been assigned, it is only the item numbers that appear in the statements of the language. The use of the numbers rather than the names within statements of the RIQS language simply saves space and time. The item name appears only on printed reports. When items from a record are reported, the name of the item is printed, directly followed by the date or value of the item.

The DATA RESTRICTIONS phase is provided so that any data which are input to the file can be checked for validity. The first group of statements in the data restrictions stage defines types for each of the items declared in record definition. The type definitions are: ALPHA-NUMERIC, ALPHABETIC, NON-NUMERIC, INTEGER, DECIMAL, AND DATE. For example:

```
DATA RESTRICTIONS
TYPES
ALPHABETIC (7)
INTEGER (3, 5)
DECIMAL (12) DATE (4)
```

Through the record definition and data restriction phase, a record structure defined for the file, and conditions of acceptance are placed on user input data. Through the INPUT DATASAVE, records of user data are actually checked and accepted as part of the file being created. The system accepts records of input data, places them in a form which allows for fast searching, and stores them as a file on a disk unit. For example:

```
INPUT DATA
( 1) C11040
( 2) CACM
( 3) 11-04
( 4) 0472
( 5) 231
( 6) THE EXPANDING WORLD OF COMPUTERS
```

- (7) HARDER E. L.
- (8) WESTINGHOUSE ELECTRIC CORP.
- (9) MANSFIELD, OHIO
- (10) THE ONWARD SWEEP OF AUTOMATIC PROCESSING
OF INFORMATION IS IMPEDED BY NINE
PRINCIPAL BARRIERS

1.3 USER INTERACTION

The on-line processor, RIQSONLINE, operates via teletype or other user terminal but is limited to file searching capabilities only. File creation, updates and indexing must be performed as batch jobs. After a RIQS file has been created and catalogued on disk, it is available for on-line interrogation.

The user begins by requesting the file that he would like to search and the system then offers to print a description of the user's file if he so desires. The file description which is printed back to the user includes the names of the items in the record as well as their structure and type. The system then asks the user to ENTER SEARCH COMMAND OR TYPE HALT after which the user proceeds to interrogate the data base. For example:

```
?IF #8'EQ "GRAPHIC" and #6 CT JAN. 1968 THEN DISPLAY  
#3 THRU #6, #8: ?END
```

The commands of a search are terminated by the word END, which initiates the searching the commands of a search.

The DISPLAY command generates reports to the teletype itself, whereas the PRINT command produces the separate report file which must be diverted to a high-speed printer.

1.4 SPECIAL FEATURES

Several useful on-line commands are available to the user to assist him in the conducting of the search.

%M-sends the arbitrary message to the computer operator at the central site. The computer operator can likewise send an answer back to the teletype.

%DROP-drops the current job and display CONTROL CARDS ? in anticipation of additional control card directives.

%STOP-stops the current job and displays CONTROL CARDS ?; additional cards then can be typed in, and they will be appended to the control card record.

%START-is used in conjunction with %STOP to reinstate job processing at the point where the job was stopped.

```
?BEFORE SEARCH PRINT "please return this output to F. Scheffler"  
BEGIN SEARCH  
IF #8 OR #13 = ("LANGUAGE OR LANGUAGES") THEN PRINT RECORD.
```

The use of the BEFORE SEARCH command allows the user some flexibility with instructions concerning his output. The HELP command is used to call in a tutorial system that will assist the user in case he gets into difficulty. The most important special feature of RIQS is the SPSS command which allows for a statistical package to be included into the system that will produce cross-tabulation tables of the variables indicated by the user.

2.0 HARDWARE

RIQS operates under the SCOPE operating system within the framework of an interactive system which was developed at Northwestern University and currently runs only on their CDC 6400. The current version of RIQS requires approximately 50,000 locations on the CDC 6400. RIQS is written primarily in Fortran IV, but it has some COMPASS assembly routines in it. Response times range from a few seconds to a few minutes for small-sized files on the order of 2,000 records.

3.0 COST

It was impossible to determine the exact cost of bringing up RIQS on a CDC 6600 at WPAFB. However, an estimate of one man-year effort to convert the system was given. The Air Force Office of Scientific Research was given listings of the system but no software implementation data were provided.

4.0 CONCLUSION

We find the following advantages with RIQS:

1. There is a tutorial mode to assist users.
2. There is on-line as well as batch capability.

3. The SPSS package is a very impressive feature, but we question whether or not AFHRL would find it useful for its particular applications.
4. Free text searching on a variety of data types is provided.

We find the following disadvantages:

1. The main disadvantage with RIQS is the fact that it is now only available on the Northwestern University machine and has been developed specifically for their software; it would be very difficult to convert to some other system.
2. The system does not permit very large data bases to be handled.
3. The system does not provide inverted file structures and therefore large data bases are not conveniently handled.
4. The supplier is not local.
5. Updating is done only in the batch or background mode. (This could be an advantage by precluding unauthorized individuals.)

Although RIQS appears to be a good system, in view of the above factors, we feel this system does not warrant further consideration.

DIALOG

1.0 DESCRIPTION

1.1 INTRODUCTION

Project DIALOG is the culmination of six years of research in information storage and retrieval conducted by the Information Sciences Laboratory of the Lockheed Palo Alto Research Laboratory. The DIALOG system offers the user an interactive command language to assist him in the retrieval of information. The current version of DIALOG has evolved from the operational experience with several large data bases, e.g., Educational Resources Information Center (ERIC), and the National Technical Information Service (NTIS).

1.2 DATA BASE CREATION

Lockheed has created several large data bases on the system and their efforts have been mainly directed in the marketing of the services provided by those data bases. DIALOG does not contain any data base creation facilities. However, keypunch card formats for the input data base could be provided and Lockheed would then be able to provide file generation programs thru the AUTO-TEXT system in order to create the desired data base that would be acceptable to DIALOG.

1.3 USER INTERACTION

DIALOG is an interactive information retrieval language and the user communicates with the system via a remote terminal. The software treats the top row of the keyboard as special function keys. Each of the major commands of DIALOG are assigned to one of those function keys. To use DIALOG, the user enters the desired commands such as BEGIN, COLLECT, DISPLAY, or PRINT, etc., by using the function keys on the keyboard of the terminal. This allows the user to enter his request quickly without typing out the full text of his request. However, this shorthand notation requires that the user know the function keys well.

The user identifies and selects those indexed terms (IT) which relate to his search topic (e.g. IT=PILOT) and build his own specialized subject heading during the search. After each entry by the user, he is informed of the number of documents which fulfill his specific search criteria. He may combine individual terms or phrases to narrow a search down to a fewer number of documents. This is done by combining the key words chosen into

a Boolean expression, which requires that all retrieved items contain the key words in the relationship desired. For example, the user may ask for the index term to be equal to sets No. 1 and No. 3 (IT=1*3), where the * is the logical connector AND. The other connectors are +, which is the logical OR, and - which is the logical NOT.

Once the desired literature references are displayed on the terminal, the user may desire to PRINT the information, or he can modify his search expression and continue the search.

1.4 SPECIAL FEATURES

The use of the EXPAND command in addition to an index term will cause a display of descriptors that are alphabetically close to the entered term. One can continue to browse through this list (or "Thesaurus") by entering the command EXPAND and the desired reference numbers.

The KEEP command allows the user to set aside selectively the items he is displaying. These items are set aside into a special category called "set 99" which can be used like any other set.

The user is also able to DISPLAY SET HISTORY during a search which causes all of the sets thus far created to be displayed.

The EXPLAIN command is available to provide an on-line supplement to the printed DIALOG USER MANUAL. Entry of this command followed by a command name or error message will provide a one page explanation.

2.0 HARDWARE/SOFTWARE

The minimum hardware configuration for an efficient use of DIALOG requires an IBM 360/40 with a 100K partition of OS. The system is able to operate in a multiprogramming environment.

The software supports several types of terminals including TTY, IBM 2741, and CRT devices. The terminal demonstrated to us was a CCI-CC30.

Communication with the Lockheed Palo Alto facility could be accomplished over the Federal Telephone System (FTS) or the Tymshare Network (TYMNET).

3.0 COST

The estimated cost of bringing up DIALOG and maintaining it for one year would be \$20,000, including the file generation programs.

It would be possible to bring up the data base on the Lockheed system for \$2,000 for the file conversion program plus \$150 per thousand records of approximately 1000 characters each in the data base. In addition there would be a storage charge of approximately \$100 per month for the data base. Connect charges would be approximately \$20/hr. to use the system after the data base were created.

The use of the ERIC files could be obtained for \$25 per connect hour plus \$.10 per item printed off-line. NTIS files are available at \$35 per connect hour in addition to the \$.10 per item printed off-line.

4.0 CONCLUSION

We find the following advantages with the DIALOG system:

1. The user's manual is well documented and easily understood.
2. It is one of the oldest and therefore proven information retrieval systems available.
3. The system provides tutorial assistance.

We find the following disadvantages:

1. The cost is high.
2. The supplier is not local.
3. A convenient means of establishing and maintaining a data base does not exist.
4. The function key concept may be "time saving" but it takes away some of the desired near-English interactive capabilities that are found in other systems.

ORBIT II

1.0 DESCRIPTION

1.1 INTRODUCTION

ORBIT II is the interactive retrieval system of System Development Corporation (SDC). It is intended primarily for the rapid retrieval of descriptive information that has been indexed. It is not intended as a general-purpose data management system.

The ORBIT II system consists of two major programs; the FILE GENERATION PROGRAM which structures the data base for storage and retrieval, and the RETRIEVAL PROGRAM, which permits users to carry out interactive searches.

1.2 DATA BASE CREATION

Each document that is to be searched is described by one-unit records stored in the system's data base file. The file must be carefully structured for retrieval and printing purposes, and this structuring task is done by the FILE GENERATION PROGRAM from data that has been prepared on 80 column card image tapes.

It is necessary to provide SDC with the specifications for the user's data base. SDC then uses these specifications to prepare a FILE GENERATION DECK, which the customer would use in conjunction with the FILE GENERATION PROGRAM. ORBIT II provides for both batch mode and on-line file updating and correction (i. e., adding, deleting, or altering unit records).

1.3 USER INTERACTION

The RETRIEVAL PROGRAM permits the user to retrieve on any designated category of information (i. e., author, title, keywords, accession number). The Boolean operators (AND, OR, NOT) may be used in any combination within a logical expression.

The EXPLAIN command allows the user to obtain an on-line explanation of any command program message. The NEIGHBOR command allows the user to retrieve index terms that are alphabetic neighbors of the search term and indicate the number of postings for each.

The PRINT command causes the program to print out information in either the on-line or off-line mode with format options.

Although ORBIT II could be used for retrieval in a batch processing mode, SDC does not recommend such use, since it does not utilize all of the special interactive features carefully designed into the system.

1.4 SPECIAL FEATURES

The system includes such aids as spelling approximation, search on truncated terms, and the printing out of indexed terms alphabetically adjacent to the specified terms.

The system also provides for ordering the outputs in terms of relevance for any one of several other numerical categories.

The system permits the use of previously stored search statements and single terms from subsequent statements through Boolean nestings. The user is also permitted to print any element of a unit record from any previous search statement that has been stored.

The TREE command is a special command used with hierarchically structured data bases. It prints the superordinate class and all the terms in the first subordinate class.

2.0 HARDWARE/SOFTWARE

ORBIT II is presently designed to operate on an IBM 360/40 and larger computers in the 360 or 370 series equipped with disk drives and teleprocessing equipment. The system operates under OS and requires a minimum of 256K bytes of core storage.

The National Library of Medicine application using SDC ORBIT II software is providing 8-10 second average response time on a mixture of simultaneously mounted data bases with an average of 36-40 simultaneous terminals operating on a 370/155.

ORBIT II is written in PL/I with some portion of the code in ALC.

3.0 COST

The ORBIT II purchase price for the multifile version is anticipated to be in the \$30,000 range. If the service were secured from SDC, at its Santa Monica 370/155, and accessed through TYMSHARE's TYMNET, then the cost would average out to about \$40 per connect hour, or \$7 to \$10 per search.

4.0 CONCLUSION

We find the following advantages with the ORBIT II system:

1. There is a tutorial mode to assist the user with problems.
2. Systems Development Corporation is a well known supplier of software.
3. This system is straightforward and simple to use.

We find the following disadvantages:

1. The cost to purchase the software is very high, especially for its limited capabilities.
2. The system is written in assembler and PL/1 and is therefore not transferrable to any system other than an IBM 360/370 without involved conversion.
3. The minimum hardware requirement is a 360/40 with 256K main memory. This appears to be very expensive.
4. The supplier is not local.

In view of the disadvantages, primarily the high costs, we do not recommend this system for implementation.

NASIS

1.0 DESCRIPTION

1.1 INTRODUCTION

The NASA Aerospace Safety Information System (NASIS) is an interactive, generalized data base management system. The system was developed for the National Aeronautics and Space Administration's Aerospace Safety Research and Data Institute (ASRDI), Lewis Research Center by Neotronics, Inc. NASA's RECON (Remote Console) system with interactive capability, served as the basis for the design of NASIS.

1.2 DATA BASE CREATION

The NASIS system is based on 2 series of file sets which are called dataplexes. A dataplex is a set of files normally consisting of an anchor file and one or more inverted files. Each file is composed of records and each record is composed of fields. There exists another file in this file set called the descriptor file, which describes the records of each of the files in the dataplex.

The entire data base for the NASIS system consists of several components: a) Field Descriptor Files, b) Linear (Bibliographic) Files, c) Inverted Index Files, and d) Thesaurus Files.

Data base security is provided to the owner of a data set by protecting the data set from other users. However, the owner of a data set may desire to share his data set or portions thereof with other users. This security is provided at both the file and the field level.

The building of the data base is done only in the batch mode. In describing the data base and the interrelationships of the files, a set of reference names are specified to identify the various levels of data within the data base. An element is the lowest defined unit of logical information. A field, the next higher level, may or may not contain multiple elements. Fields are associated to form records. Each uniquely defined data set is composed of multiple keyed records.

Each source file which is to be loaded into a NASIS dataplex consists of records with uniquely defined fields. The person responsible for loading the file first defines to the system how he would like the resultant dataplex to be formatted. The actual loading process reads the source file and writes the dataplex.

1.3 USER INTERACTION

Interaction with the NASIS system for the retrieval of information is available in both batch and on-line modes. On-line use via a remote terminal is the normal manner by which users interrogate the data base. After dialing into the system, the user indicates that he would like to BEGIN NASIS after which a NASIS ID and a PASSWORD are requested by the system to ensure the security of the NASIS files. Both the valid NASIS ID and the valid PASSWORD are unique to each user and must be obtained prior to operating the system. If the system cannot recognize both codes as valid, the user cannot continue. After the user has logged onto the system, he enters the NASIS command RETRIEVAL, which causes the system to print out the names of all files that are available to the particular user for retrieval. The user then selects one of those files that he would like to work with and enters its name when prompted by the system.

The user then has several options from which he may select to proceed in his browsing of material. One such option is to enter the command FIELD. This command causes the system to list all of the field names which have previously been defined in the file that he has just selected.

Another alternative available to the user is the command EXPAND. This command causes the display of those terms which alphabetically surround the term from the cross reference files specified in the inverted index. The number of terms displayed depends on the number of lines on the terminal screen. To display the next following set of terms on the screen, the user must enter the PAGE command. Looking backward through the index can be achieved by paging backwards.

In the retrieval process, the system asks the user to enter the index field name for the term that the user has requested to be expanded. After the user responds, the system prints a list of the terms and also supplies a line number associated with each term. The user then selects logical combinations of line numbers to which the system assigns a set number for the current search strategy. The command SELECT followed by the logically combined line numbers initiates the retrieval process. The binary Boolean operators of +(OR), *(AND) or -(NOT) may be used with the SELECT expression. The user may select combinations of the previously determined SELECTS which are identified by set numbers and then display the information according to some predetermined format that he would like. A default feature causes the output to be formatted according to a standard format in case the user does not actually specify the format.

1.4 SPECIAL FEATURES

The SAVE command is used to save the current screen image appearing in the output data of the terminal display screen. This information is stored in another special set and the user can at any time request output of the contents of this set.

The user can use the RESTART command to restore the retrieval system to the point in his strategy that was being executed either when the system crashed or when he was forced off the system. The command strings comprising the then-current strategy are saved. Each command string is retrieved and re-executed individually until the 'save current strategy' is exhausted; the user can then continue his strategy.

The NASIS system provides a means of generating reports which allows the user to format column headings, report titles, report pagination, arithmetic sums for any field name, and the numeric tally for any field name. This special feature is called the report generator and allows the user to produce an overall report in a very flexible manner.

The statistical capability provided within the NASIS system serves multiple purposes. Retrieval statistics consist of the connect time and central processing unit (cpu) time, the total strategy length, the number of strategies currently stored, the data set names of the stored strategies, the total number of terminal sessions and the date of the first and last terminal session.

The retrieval of statistics consists of two reports. The first report contains the activity of the EXPLEX, the average length of the strategy per sessions, the total connect and the total cpu time, the number of strategies currently stored, the average number of times COMMAND was invoked per session and the total number of terminal sessions. The second report contains the total number of transactions per maintenance run, the total records currently in the dataplex anchor file, the average percentage of the file affected by the maintenance runs, the frequency of maintenance runs, and statistics per maintenance run of add, delete and update transactions.

The most significant special feature of the NASIS system is that it is a highly sophisticated data base manipulation language which encompasses the realm of information storage and retrieval as well as management information systems. Therefore it has great potential for many general purpose functions. The system is more sophisticated than is required for the envisioned AFHRL applications that we are investigating.

2.0 HARDWARE

The NASIS system currently operates on an IBM 360/67 under the Time Sharing System (TSS). However, a conversion of the system is being made to operate on any 360/370 under OS by January 1, 1973.

The system currently supports the CC-30 and IBM-2741's as remote terminals to the system. At this time it is estimated that the present conversion to OS will require one megabyte of core. It is being written in PL/1 and assembly language and will support 20 simultaneous users.

3.0 COST

The use of the IBM 360/67 system at the Lewis Research Center by AFHRL is not possible unless a policy change is made. However, the Food and Drug Administration in Washington could be checked regarding the availability of their IBM 360/67 system. There would be no cost for the OS software for the NASIS system to any other Government agency. However, we estimate a three man month effort for a consultant from Neoterics would be necessary to assist in the implementation of the system on any IBM 360/370 computer for AFHRL.

4.0 CONCLUSION

We find the following advantages with the NASIS system:

1. The software is available at no cost and would require minimal cost to install on an IBM 360/370.
2. It is a comprehensive system which includes facilities beyond information storage and retrieval. (i. e., the data base management system)
3. Neoterics is a competent supplier.
4. The system has on-line and batch capability.
5. There is a tutorial mode to assist the user.
6. Since the system is designed according to the CODASYL Committee data base management system standards, any existing files could easily be transported to the NASIS system.

We find the following disadvantages:

1. The system is written in PL/1 and therefore is not transportable to any system other than an IBM 360/370.

2. To build a data base, a trained PL/I programmer is required to get the job done.
3. The system may be too comprehensive and therefore the cost to operate it may be high. The system might be somewhat cumbersome to operate.
4. Updating is done only in batch mode. (This could be an advantage by precluding unauthorized individuals.)

QUERY/UPDATE

1.0 DESCRIPTION

1.1 INTRODUCTION

QUERY/UPDATE (QU) is an interactive, generalized data base management system. The system is currently under development by Control Data Corporation.

QU is a computer program to interrogate and maintain mass storage data files. Program control directives are submitted through remote terminals or in punched card format. The personnel using QU are expected to be under the guidance of an individual familiar with computer operations; his responsibility is to serve as a data administrator. The QU user requires minimal computer orientation to submit directives that will cause the program to list requested information, compare data, remove and insert entries or select and modify portions of the data content.

1.2 DATA BASE CREATION

The description of a data base is known as a SCHEMA. The data administrator names a SCHEMA and determines each file to be included as an AREA within the named SCHEMA. Each type of information related to an AREA is described as a RECORD. Data within a RECORD is grouped logically and is given descriptive group or elementary data names. For instance, a RECORD could contain a group data name such as ADDRESS. This group data name would be divided further to contain elementary data names for STREET NUMBER, STREET, CITY, STATE, AND ZIP CODE.

Following the description of data, actual information is placed in the mass storage files according to the organization determined by the data administrator. Unauthorized access to established mass storage filing information is prevented through the use of security codes. Passwords are given to the users for terminal entry along with any other additional security identification codes needed for data bases.

The data base organization has been designed according to the CODASYL Committee on Data Base Management Systems standards.

A DATA Definition Language (DDL) and Data Manipulation Language (DML) are currently being defined by Control Data to fit the standards set by this committee.

1.3 USER INTERACTION

A system program named INTERCOM connects user terminals to the computer. This program directs the flow of information between terminal users and programs that share computer time. INTERCOM permits time sharing use of programs other than QU in a multiprogramming mode.

The QU program receives directives, performs data base maintenance operations, and produces listings or displays results when operated interactively from a terminal or noninteractively in the batch mode. Tutorial information concerning QU operation can be requested by typing the single word HELP. Tutorial material can be requested from the INTERCOM program by typing TEACH. In addition to the tutorial displays for the user, QU responds to an EXHIBIT directive and successively displays at the terminal: the AREA-NAME, RECORD-NAME, the SEQUENCE-KEY-DATA-NAME, and the ORGANIZATIONAL-DATA-NAMES of information groups. These may also be used in the batch mode.

For on-line retrieval and display operations, a set of commands is available. The USE directive identifies the file to QU that the user desires to process. At this time the user also gives any permanent file password necessary to allow him to perform operations on the file. The EXHIBIT directive displays the AREA-NAME and the DATA-NAME to show organization of the information. The DISPLAY directive retrieves specified information and extracts the named fields for display. This information may be displayed upon the user's terminal or written to a designated file which the user can call up later. The IF directive indicates the conditions that QU should satisfy when selecting records for subsequent directives. Any of the six usual relational operators may be used in the conditional expression. The connective Boolean operators AND, OR, XOR may also be used. XOR, an exclusional OR, requires either one condition or the other to be present but not both simultaneously.

On-line updating of records can be performed. The DELETE directive removes a specified record from the file, while the INSERT directive creates a record entry and inserts it into an area. Another option available is the UPDATE directive which notifies the system that information is to be modified in one or more records. The MOVE directive places a literal value in a data field after making any format revisions required by the specifications of that data field. The STOP directive causes QU to close active files and return the permanent files to the operating system. Control is then returned to INTERCOM for general processing.

2.0 HARDWARE/SOFTWARE

QU is designed for the Control Data CYBER 70/Models 72, 73 and 74 and 6000 Series Computer Systems. The control system program named INTERCOM is required in order to permit time sharing use of the system and to provide terminal access. The amount of core requirements to use QU have not yet been determined since the software is currently under development.

3.0 COST

Cost information is also not available at this time since the software is still under development. It is expected however that the cost will be on the order of a few thousand dollars per month rental for the hardware and software.

4.0 CONCLUSION

We find the following advantages with the QU system:

1. It is a very sophisticated system.
2. Control Data is a well known supplier of hardware and software.
3. The system has on-line and batch capability.
4. There is a tutorial mode to assist the user with problems.
5. There are few directives, which makes it simple for those people searching a data base to perform the activities.

We find the following disadvantages:

1. The software will most likely not be transferrable to any other system.
2. The minimum hardware requirement will most likely be somewhat expensive.
3. This system encompasses more than just an information storage and retrieval system, and therefore may be more sophisticated than necessary.
4. Programmers will be required to build the data base.

The QU system provides many capabilities that we would not expect AFHRL to use at this time; however, this certainly is a system that should be considered again.

STAIRS

1.0 DESCRIPTION

1.1 INTRODUCTION

The IBM S storage and Information Retrieval System (STAIRS) consists of a program package that enables the user to search documents through dialogue with the system, to produce or display documents on various output devices, to create data bases by automatically processing text (without manual indexing of documents), and to extend and maintain data bases by means of an updating utility program.

STAIRS provides a broad spectrum of search capabilities as well as flexible and varied output options for retrieved documents.

1.2 DATA BASE CREATION

A document is a body of writing that is mostly textual and has a uniquely identified document number assigned externally by the user. A document is structured into information elements that are made up of one or more "paragraphs". The document "paragraph" comprises anything from an actual paragraph of text, to a single word, to a formatted data field. Each paragraph is tagged with a three character alphanumeric "paragraph code".

When building a STAIRS data base, the basic unit of input used is this "paragraph record" which represents one paragraph of text for a document. All paragraph records which contain the same twelve character document number comprise one full document.

A set of programs called TEXT-PAC provides for the user a method of capturing, correcting, editing, indexing and checking documents that he wishes to put into the data base in the specifications required by STAIRS. Each document must have an identification number. Within a document, paragraphs should be identified. For example, there may be a title, a source, author or authors, abstract or other text, and additional paragraphs as needed. TEXT-PAC can accept and process text in upper/lower case. It is possible to capture data at the source using MT/ST tapes and to convert the data into TEXT-PAC format.

The output from the TEXT-PAC program is submitted to the Text File Build Program which transfers the text to the Text data set (TEXT). In addition, the Text Index data set (TNDX) is generated, which points to the

beginning of each document in the Text data set, identifies the document by number, and contains, in formatted fields, text extracted from the document. Words or paragraphs that are considered insignificant are referred to as stop words or stop paragraphs. They are furnished to the program through the STOP data sets.

For each significant word, the Inverted data set requires a list of its occurrences. The Text File Build program generates the Inverted data set and the Dictionary data set and puts the output onto disk storage.

In summary, for each data base in the system, four files are automatically created: a Dictionary file, which contains every unique significant word (i. e., words other than "and, " "the, " etc.) in that data base, along with synonyms, the number of occurrences, and the number of documents; an Inverted file, containing pointers (document number, paragraph number, sentence number, and word number) to every occurrence of that word; a pointer to the Text Index file, which contains control and security information, and a pointer to the text; the Text file itself.

STAIRS data bases can be updated off-line by means of the Data Base Merge program. Up to four data bases can be merged by combining corresponding data sets and reorganizing them. A special feature of this facility provides for the removal of documents in each of the data bases. This group of data base utility programs that accepts the output of TEXT-PAC is known as A Query and Retrieval Interactive Utility System (AQUARIUS). AQUARIUS precludes unauthorized access to data bases. Each of the data bases has its privacy to authorized users protected at the data base, and at document, paragraph and field levels.

1.3 USER INTERACTION

STAIRS is a user-oriented information storage and retrieval system that permits any authorized user to engage in a step by step dialogue with the computer until the relevant information is found. Interaction with the STAIRS system for the retrieval of information is available in both batch and on-line mode.

The user can conduct a search for documents that have, or do not have certain keywords. In his specification for search criteria, the user can link keywords with the logical operators AND, OR, NOT, thus creating a highly specific search if he so desires. He can require that the specified keywords be in the same paragraph, the same sentence, or that they be adjacent in a word phrase in order to satisfy his search criteria. The user can cause retrieved documents to be printed, and he can save a set of queries for later searches. He can, with one query, instruct the system to search as many as 16 different data bases.

1.4 SPECIAL FEATURES

The HELP function prompts the user whenever he is unsure of functions or command format. This function provides tutorial information when the user signs on to the system. If the user requests HELP at this time the system provides a description of the available functions and commands. The user is prompted to ask for more information or to proceed under normal mode.

During processing, the user can invoke the HELP function in any situation. The HELP processor shows the user the area in which he requires assistance.

2.0 HARDWARE/SOFTWARE

Most of the STAIRS program modules are written in Assembler language. Some are written in PL/1. All programs must be compiled and executed under the IBM Operating System. The data management facilities used for execution of the on-line dialogue system AQUARIUS are provided by the IBM Customer Information Control System (CICS). The STAIRS system currently operates only on an IBM system 360 or system 370.

3.0 COST

Under IBM's unbundling policies, the license fee for CICS is \$700 per month, for STAIRS is \$500 per month, thus making a total of \$1200 per month to use the software.

4.0 CONCLUSION

We find the following advantages with the STAIRS system:

1. It is a very sophisticated system.
2. It is highly user oriented.
3. IBM is a well known supplier of hardware and software.
4. The system has on-line and batch capability.
5. There is a tutorial mode to assist the user with problems.
6. The MT/ST offers a convenient means of inputting documents with TEXT-PAC programs.

We find the following disadvantages:

1. The cost to use the software is very high regardless of what system it is implemented on.

2. The system is written in assembler and PL/1 and is therefore not transferrable to any system other than an IBM 360/370 without conversion.
3. It must run under CICS, also making it not transferrable.
4. The minimum hardware requirement is a 360/40 with 256K main memory. This appears to be expensive.

The STAIRS system provides many capabilities that AFHRL would probably not use at this time; however, this certainly is a system that should be considered again.

APPENDIX B

USER'S MANUAL FOR THE MODEL DATA
BASE OF COMPREHENSIVE ABSTRACTS
USING AVIONICS CENTRAL